

UNIVERSIDADE FEDERAL DO PARANÁ

LUIZ RENATO MARTINI FILHO

A HYBRID ARIMA AND ANN MODEL FOR SYNTHETIC STREAMFLOW  
GENERATION

CURITIBA

2018

LUIZ RENATO MARTINI FILHO

A HYBRID ARIMA AND ANN MODEL FOR SYNTHETIC STREAMFLOW  
GENERATION

Dissertação apresentada ao Programa de Pós Graduação em Engenharia de Recursos Hídricos e Ambiental, Área de Concentração em Engenharia de Recursos Hídricos, Departamento de Hidráulica e Saneamento, Setor de Tecnologia, Universidade Federal do Paraná, como parte das exigências para a obtenção do título de Mestre em Engenharia de Recursos Hídricos.

Supervisor: Prof. Dr. Daniel Henrique Marco Detzel

CURITIBA

2018

Catálogo na Fonte: Sistema de Bibliotecas,  
UFPR Biblioteca de Ciência e Tecnologia

F481h

Martini Filho, Luiz Renato

A hybrid ARIMA and ANN model for synthetic streamflow generation  
[recurso eletrônico] / Luiz Renato Martini Filho. – Curitiba, 2018.

Dissertação - Universidade Federal do Paraná, Setor de Tecnologia,  
Programa de Pós-Graduação em Recursos Hídricos e Engenharia Ambiental,  
2018.

Orientador: Daniel Henrique Marco Detzel.

1. Hidrologia. 2. Usinas hidrelétricas. 3. Estações fluviométricas. I.  
Universidade Federal do Paraná. II. Detzel, Daniel Henrique Marco. III. Título.

CDD: 627

Bibliotecária: Vanusa Maciel CRB- 9/1928

**TERMO DE APROVAÇÃO**

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ENGENHARIA DE RECURSOS HÍDRICOS E AMBIENTAL da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **LUIZ RENATO MARTINI FILHO**, intitulada: "**A HYBRID ARIMA AND ARTIFICIAL NEURAL NETWORK MODEL FOR SYNTHETIC STREAMFLOW GENERATION**", após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de Mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 22 de Março de 2019.



DANIEL HENRIQUE MARCO DETZEL  
Presidente da Banca Examinadora



ANA PAULA OENING  
Avaliador Externo (LACTEC)



ELOY KAVISKI  
Avaliador Interno (UFPR)



MARCELO RODRIGUES BESSA  
Avaliador Interno (UFPR)

*Aos meus pais, Luiz Renato e Ana Rita, às minhas  
irmãs, Maira e Renata, aos meus sobrinhos,  
Victoria e Anthony e à minha namorada Poliane,  
pelo carinho, amor e suporte.*

## **AGRADECIMENTOS**

Esse trabalho não seria possível sem a ajuda de terceiros a qual veio de várias formas distintas, com suporte acadêmico, financeiro e emocional.

A primeira pessoa a agradecer é o meu orientador e professor, Dr. Daniel Henrique Marco Detzel, cuja ajuda foi além do que lhe era esperado. Além do suporte técnico e intelectual agradeço também por ter compreendido todas as dificuldades e ansiedades que surgiam pelo caminho e pelos momentos de descontração, que transformaram a relação de aluno e professor em uma amizade levada para além do ambiente acadêmico.

Agradeço aos membros da banca, Prof. Ana Paula Oening, Prof. Eloy Kaviski e Prof. Marcelo Rodrigues Bessa, juntamente à Prof. Dra. Miriam Rita Moro pelas recomendações e contribuições que certamente agregaram um grande valor intelectual ao trabalho.

Ao Programa de Pós Graduação em Engenharia de Recursos Hídricos e Ambiental juntamente com todos os seus professores e funcionários por toda a ajuda prestada, em especial ao Prof. Cristóvão Vicente Scapulatempo Fernandes pelo incentivo que me foi de grande importância.

Aos colegas bolsistas e estagiários, pelo ambiente de trabalho sadio e pelas longas discussões que extrapolavam a temática do trabalho e por vezes a esfera científica, mas que certamente ampliaram meu horizonte e mudaram minha visão de mundo.

Ao Instituto de Pesquisa Lactec e à Copel pelo suporte financeiro que veio na forma de uma bolsa de mestrado junto ao projeto de P&D Lynx.

Nem toda a ajuda é de cunho acadêmico e tão importante quanto o suporte intelectual é o suporte emocional. Agradeço a meus pais Luiz Renato e Ana Rita pela criação que me foi dada com muito amor e carinho, por me mostrarem o que é certo e o que é errado. Por me ensinarem como ser uma pessoa digna. Por me mostrarem que primeiro vem a obrigação e depois a diversão. Por terem sempre orgulho de mim, mesmo quando eu não tinha. Por se fazerem sempre presentes sempre em minha vida. Pelos abraços felizes em todas as minhas conquistas, pelos abraços de conforto nos momentos de dificuldade. Por se desprenderem de suas próprias vontades para tornar os

sonhos meus e de minhas irmãs possíveis. Sou eternamente grato não só por esse trabalho, mas por me darem a vida e fazerem de tudo para que ela seja a melhor possível.

À minha irmã Maira, que sempre esteve ali para mim, mesmo que por vezes eu não percebesse. Pelas conversas e cervejas, pelos momentos de alegria e de tristeza e por todas as vezes que me mostrou a importância da família.

À minha irmã Renata e aos meus sobrinhos Victoria e Anthony, que mesmo longe fisicamente me incentivam muito mais que eles imaginam. Pelas conversas por Skype que muitas vezes trouxeram luz e cor aos dias mais cinzas.

À minha namorada Poliane, a qual nunca saiu do meu lado. Em todas as dificuldades ela estava lá para me mostrar que não era tão difícil assim e que eu era capaz. Obrigado Poliane por toda a paciência, companheirismo amor e carinho e por acreditar em mim.

Aos meus amigos de longa data Jhonatan, Mateus, Vitor e Rafael, pelos momentos de descontração que ajudaram a manter o foco nos momentos de estudo. As risadas e cervejas foram um dos pilares que me deram suporte nesses dois anos.

A João Henrique, Matheus e Bruno, amigos feitos durante o mestrado, por ajudarem a tornar as semanas mais leves e os finais de semana mais divertidos, além de conversas técnicas que me guiaram em momentos de dificuldade.

Por fim agradeço a Deus, pois sem Seu amor e amparo eu nada seria!

**Short cuts make long delays!**

*J.R.R. Tolkien*



## RESUMO

Estudos das características das séries de vazão são de grande importância no campo da hidrologia, representando uma ferramenta útil para a previsão de cheias, mitigação de danos em catástrofes, projeto e operação de reservatórios, geração hidroelétrica, projetos de barragens e vertedouros, dentre outros. Os dados históricos nem sempre estão completos ou corretos e isso tem efeito direto na confiabilidade dos resultados. Séries sintéticas são consideradas um excelente dispositivo de extrapolação para a solução de problemas complexos. Nesse sentido, o grupo de modelos *Box & Jenkins* (i.e., ARIMA) vem sendo usado com esse propósito por décadas, destacando-se por sua capacidade de calcular e replicar a estrutura de persistência da série histórica. Contudo, esses modelos são lineares, ao passo que o comportamento dos corpos hídricos é não-linear. As redes neurais artificiais (i.e., ANN) vêm como alternativas não-lineares a este grupo de modelos lineares clássicos. Esta dissertação propõe o acoplamento entre os modelos ANN e ARIMA para a geração de séries sintéticas de vazão. O objetivo do acoplamento é melhor computar as estruturas de persistência e não-linearidade juntando um modelo estocástico linear com um modelo não-linear do tipo “*black-box*”. O modelo foi testado para seis estações fluviométricas do rio Iguaçu, na região sul do Brasil. Em seguida, estatísticas de longo e curto termo foram usadas para verificar a adequação do modelo para a geração de séries sintéticas de vazão. Uma análise comparativa foi feita considerando um modelo ARIMA tradicional ajustado às mesmas estações. Por fim, os dois modelos reproduziram com sucesso as estatísticas históricas, contudo, o modelo híbrido foi superior na preservação do coeficiente de assimetria e das vazões mínimas.

Palavras-chave: Hidrologia Estocástica. Séries Sintéticas de Vazão. ARIMA. Redes Perceptron Multicamadas. Modelo Híbrido.

## **ABSTRACT**

Streamflow characteristics studies are of great importance in the hydrology field, representing a resourceful tool in procedures such as flood forecasting, damages mitigation in catastrophes, reservoir design and operation, hydroelectricity generation, dam and spillway design, among others. The recorded data are not usually complete or corrected; it can damage the study reliability. The synthetic series generation is suggested as a resourceful extrapolation device for the solution of complex problems. Thus, the Box & Jenkins (i.e. ARIMA) models are being used on this purpose for decades and are especially good in computing and replicating the persistence structure of the historical. However, these models are linear, whereas catchments behaviors are non-linear. The artificial neural network (ANN) models come as non-linear alternatives to the classic linear ensemble of models. This master thesis proposed a coupling between an ANN and an ARIMA model for synthetic streamflow series generation. The purpose is to better address both persistence and non-linear patterns by joining a stochastic linear and a black-box non-linear models. The model was performed for six gauging stations within the Iguaçu river, on the South region of Brazil. Furthermore, long- and short-term statistics are used to verify the adequacy of the model for synthetic streamflow generation. A comparative analysis considering a single ARIMA model at the same condition. Finally, both models successfully reproduced the historical statistics, the hybrid model, however, better preserved the skewness and streamflow minimum values.

Key-words: Stochastic Hydrology. Synthetic Streamflow Series. ARIMA. Multilayer Perceptron. Hybrid Model.

## LIST OF ILLUSTRATIONS

Figure 1. Differentiation between a stochastic process realization and the historical series.....	5
Figure 2. Synthetic series generation (left) and series forecasting (right).....	6
Figure 3. Stages in the iterative approach to model building. (Font: Box et. Al., 2008) .....	14
Figure 4. General architecture of a two hidden layers ANN .....	19
Figure 5. Architecture of a Multilayer Perceptron with one hidden layer.....	21
Figure 6. ARMA-ANN coupling scheme .....	26
Figure 7. MLP tested architectures.....	27
Figure 8. Map of the Parana River basin.....	29
Figure 9. Map of the Iguaçu River basin, divided in Upper, Intermediate and Lower Iguaçu.....	30
Figure 10. Map of the Iguaçu Basin.....	32
Figure 11. Historical and AR(1) residual series at Foz do Areia .....	36
Figure 12. ACF and PACF at Foz do Areia .....	37
Figure 13. Expected and estimated outputs at Foz do Areia .....	39
Figure 14. Training Regression .....	40
Figure 15. Synthetic series of residuals – Foz do Areia .....	41
Figure 16. Synthetic streamflow series – Foz do Areia 1 .....	41
Figure 17. Synthetic streamflow series – Foz do Areia 2 .....	42
Figure 18. Monthly statistics – Foz do Areia. Bars indicate means, and lines indicate standard deviations.....	44
Figure 19. Autocorrelation Functions – Foz do Areia .....	44
Figure 20. ACF and PACF at Foz do Areia .....	55
Figure 21. ACF and PACF at Segredo .....	55
Figure 22. ACF and PACF at Salto Santiago .....	55
Figure 23. ACF and PACF at Salto Osório .....	56
Figure 24. ACF and PACF at Gov. José Richa .....	56
Figure 25. ACF and PACF at Baixo Iguaçu .....	56
Figure 26. Residual series histograms .....	57
Figure 27. Monthly statistics – Foz do Areia.....	58
Figure 28. Monthly statistics – Segredo. ....	58

Figure 29. Monthly statistics – Salto Santiago.....	58
Figure 30. Monthly statistics – Salto Osório. ....	59
Figure 31. Monthly statistics – Gov. José Richa.....	59
Figure 32. Monthly statistics – Baixo Iguaçu. ....	59
Figure 33. Autocorrelation Functions – Foz do Areia .....	60

## LIST OF TABLES

Table 1. ACF and PACF theoretical behavior.....	16
Table 2. Hydroelectric plants data .....	33
Table 3. Descriptive statistics .....	34
Table 4. Results of the trend tests.....	35
Table 5. BIC results for the ARIMA models tested .....	37
Table 6. ARIMA (1,0,0) parameters and theoretical validation p-values.....	38
Table 7. Optimal architectures and RMSE .....	39
Table 8. Short-term statistics and uncertainty .....	43
Table 9. Long-term statistics and uncertainty .....	45

## LIST OF ABBREVIATIONS

ACF	-	Autocorrelation Function
ANA	-	National Water Agency of Brazil
ANN	-	Artificial Neural Network
AR	-	Autoregressive Integrated Moving Average
ARIMA	-	Autoregressive Integrated Moving Average
ARMA	-	Autoregressive Moving Average
BIC	-	Bayesian Information Criterion
CV	-	Coefficient of Variation
ENSO	-	El Niño Southern Oscillation
GRNN	-	General Regression Neural Network
MA	-	Moving Average
MLP	-	Multilayer Perceptron
NN	-	Neural Network
ONS	-	Brazilian Operator of the National Electricity System
PACF	-	Partial Autocorrelation Function
PARMA	-	Periodic Autoregressive Moving Average
RBF	-	Radial Basis Function
RMSE	-	Root Mean Square Error
SARIMA	-	Seasonal Autoregressive Moving Average
SVM	-	Support Vector Machine

## TABLE OF CONTENTS

INTRODUCTION .....	1
1        Theoretical Background .....	4
1.1       Hydrologic Modelling .....	4
1.2       Synthetic Hydrology .....	7
1.3       Considered Models .....	10
1.3.1       Box & Jenkins Models .....	11
1.3.2       Artificial Neural Network Models.....	18
2       Hybrid Models and Proposed Scheme .....	24
3       Study area .....	29
3.1       Characteristics .....	30
3.2       Climate .....	31
3.3       Selected stations.....	31
4       Results and discussion .....	35
4.1       Preliminary Analysis: Stationarity Comparison.....	35
4.2       ARIMA model .....	36
4.3       ARIMA-ANN model .....	38
4.4       Comparative analysis .....	42
5       Conclusion .....	46
REFERENCES .....	48
APPENDIX .....	55
A.1       Autocorrelation and partial autocorrelation functions .....	55
A.2       Histogram plots - Residuals .....	57
A.3       Monthly average and standard deviation .....	58
A.4       Autocorrelation function comparison .....	60

## INTRODUCTION

Ever since the dawn of civilization, mankind has been dependent on water, and as the humanity evolved, the technologies regarding the water resources evolved along with it. From ancient human tribes, that would settle near water streams mainly to facilitate access to drinking water and fishing, until the modern society that rely on water for so many purposes besides drinking and fishing, such as irrigation, transportation, hydroelectricity, recreation, among others.

With an increase in water usage, the need for a better understanding of river regimes behavior has emerged. The complexity regarding the hydrological phenomena led to the development of mathematical tools for hydrological modeling, turning hydrology into a more feasible science. At the beginning of the last century, synthetic hydrology took place and brought a new set of models, largely being used in water resources related studies until nowadays.

Among the many models for synthetic streamflow generation, a group of models that stands out is the Box & Jenkins set of models, also known as autoregressive integrated moving average (ARIMA) models. These models have been used in the past few decades, and their success is mainly due to their capability of addressing the persistence of a time series. Nevertheless, this established ensemble of models have a linear formulation, whilst hydrological events mostly present non-linear behavior. Alternatively, the non-linear artificial neural network based models, also called ANN models, are capable to compute those non-linear patterns and demonstrate a good performance, notwithstanding the inability to evaluate the persistence. In fact, both ARIMA and ANN models had shown good results over the years and are vastly used in water resources modeling nowadays. However, none of them has a comprehensive approach regarding persistence and non-linearity issues simultaneously. Thus, a hybrid model may successfully address both issues and perform better than a single ARIMA model.

This research project aims to verify the adequacy of a hybrid model between ARIMA and ANN for synthetic streamflow generation in comparison with the classic ARIMA modelling. For the comparison, a study case at the Iguaçu basin for six catchments at the six hydroelectric plants within the Iguaçu River is



performed by the two models with further comparison of the results. Hence, the general objective of this dissertation is to verify whether the hybridization of an ARIMA model with an ANN based model for synthetic streamflow generation improves the results in relation to the synthetic series generated by a single ARIMA model. The Box & Jenkins models are especially efficient in computing the persistence of a series. However, its linear equation cannot address the non-linearity of a time series. Moreover, to avoid numerical issues regarding the synthetic series generation, the integration portion of the model must be suppressed, providing a stationary model (SALAS et. al., 1980).

Aiming to fulfill these gaps, while maintaining the serial correlation approach given by the ARIMA formulation, the proposal is to couple this consolidated model with an ANN based model. In order to achieve that, the following specific objectives must be met:

- (i) Address the non-linearity and non-stationarity issues concerning the ARIMA model;
- (ii) Implement and train an ANN model capable of generating synthetic series;
- (iii) Couple the ANN to the ARIMA model;
- (iv) Generate synthetic series using a single ARIMA model, for comparison with the hybrid model.

In the following section a theoretical background regarding hydrologic modelling, synthetic hydrology and the considered models is given. The section starts with a differentiation between deterministic and stochastic approaches, as well as a short definition of empirical models, followed by detailing synthetic hydrology and ending with the theories behind ARIMA and ANN models. The second section addresses hybrid models and follows to present the hybrid model proposal for this research, detailing the model coupling and working. The third section details the study area, followed by a forth section presenting and discussing the results, starting with the preliminary results regarding the non-stationarity issue, followed by the single and hybrid model specific results and ending with a comparative analysis between them. Section five presents the conclusions for this study and recommendations for future research. The

Appendix section presents a comprehensive list of graphs with the results for all the six gauging stations.

## 1 THEORETICAL BACKGROUND

For water resources planning and management, streamflow modeling and forecasting play an important role both for short- and long-term temporal scales. The former is needed for flood analysis and hazard mitigation systems, while the latter is essential in operation and planning of reservoirs, hydropower generation, among others (YASEEN et al., 2015).

### 1.1 HYDROLOGIC MODELLING

Streamflow data are the main objects of study in the water resources management. Studies on flood forecasting, city planning, operation of reservoirs, water distribution, water quality, among others depend directly on the river discharges. The uncertainty regarding hydrological processes explains the use of statistical and stochastic principles in hydrological modelling. Statistical analyses are taken using historical data, which may be short in length, incomplete or incorrect, weakening the significance of results. In addition to that, Matalas (1967) and Jackson (1975) affirm that the occurrence of the same historical series in the future is improbable and the worst recorded flood in a catchment is unlikely to be the worst possible flood for that basin. In response to these, many hydrologists use synthetic streamflow generation in order to increase robustness of data.

The need for reliable information lead to the usage of several mathematical models in order to better understand and describe the hydrological phenomena, accordingly these models can be classified as deterministic or stochastic models (HIPEL and McLEOD, 1994). In essence, those models that consider the probability of occurrence of an event are classified as stochastic, whilst models not considering it are deterministic.

For deterministic models, the randomness of the variables is not taken into account and the process relies on laws other than the statistical. Thus, a deterministic model always produces the same output for a given input under the same initial conditions. Additionally, deterministic models are mainly physically-based, meaning that those models rely on geometrical characteristics and physical phenomena inherent to the process and are ruled by the laws of physics (CHOW, 1964; DOOGE, 1973).

The implementation of such models often involves a large number of variables, hence incurring in some challenging issues regarding computational-cost, field measurements, and the determination of the relevant physical parameters (SALAS et al., 1980; HIPEL, K. W. and McLEOD, 1994 ; KASIVISWANATHAN et al., 2016; MARTINI FILHO et al., 2017).

Alternatively, data-driven models which rely on historical observations and are able to describe structures evolving over time through a probabilistic approach are considered stochastic. The probability structure regarding the evolution of a process over time characterizes a stochastic process, meaning that streamflow series are, by definition, stochastic phenomena. Those processes can be linear or non-linear, moreover, when related to hydrology, they frequently present both linear and nonlinear portions.

The stochastic hydrology considers that there is an infinite number of possible realizations within the same stochastic process and understands that, among them, the historical series is the sample that was registered (Figure 1). On the other hand, synthetic series are artificially generated series, representing alternative scenarios to the registered series.

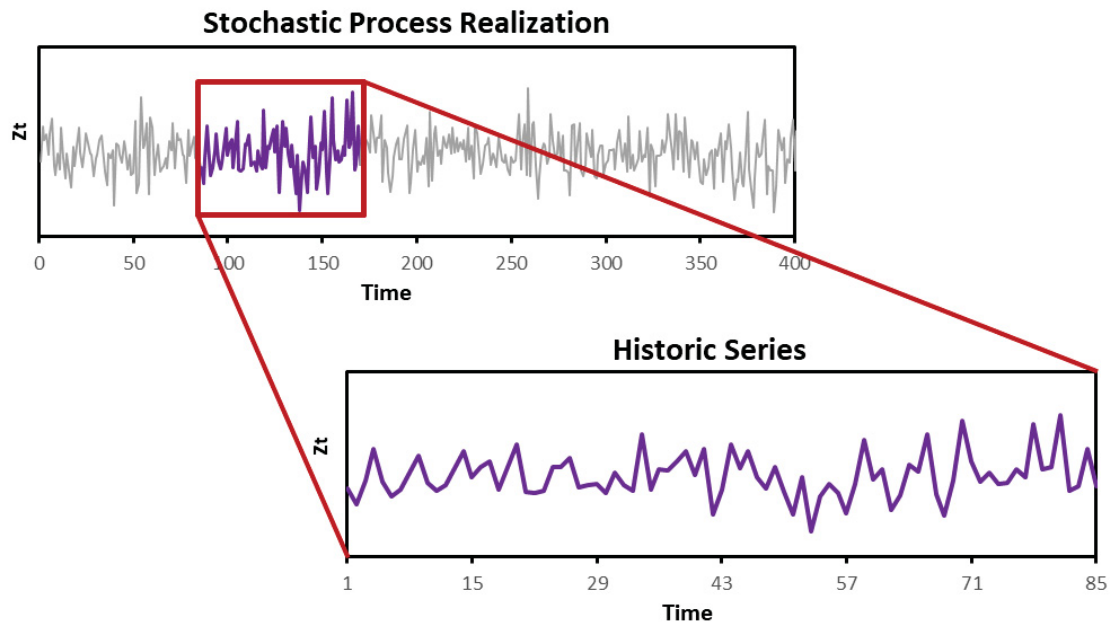


Figure 1. Differentiation between a stochastic process realization and the historical series

Herein, clarify the difference between synthetic generation and series forecasting is crucial. The former refers to the generation of several scenarios equally probable to the original. These scenarios lack temporal reference, being impossible to specify the day and year a series starts. On the other hand,

temporal reference must be considered when aiming to forecast a series, since the concern is the future evolution of a previously known event. Figure 2 illustrates that distinction for a variable  $Z_t$ . Therefore, it is relevant to hydrological studies to complement the historical data with synthetic streamflow in order to better represent the stochastic process for planning, design and operation of various water resources systems. Additionally, stochastic hydrology has the advantage of generating synthetic streamflow sequences which are statistically related to the registered series. That, based on the premise that the catchments future behavior will be similar to the one registered, makes it possible to produce many feasible scenarios for the catchment. (SALAS et al., 1980; AHMED and SARMA, 2007; HIPEL, K. W. and McLEOD, 1994; TAGHI SATTARI et al., 2012).

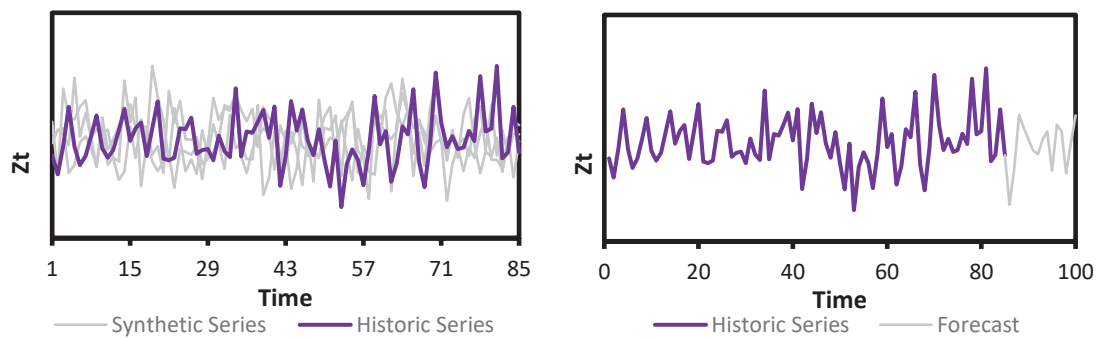


Figure 2. Synthetic series generation (left) and series forecasting (right)

Mainly, stochastic models, similarly to stochastic processes, are classified as: (i) Linear (i.e. ARIMA models; disaggregation models); (ii) Non-linear (i.e. fractional Gaussian noise models; neural network-based models). The former and the latter demonstrate good performances in hydrological studies, despite the fact that the natural processes usually present both linear and non-linear portions and these models can address one pattern only (MARTINI FILHO et al., 2017; OCHOA-RIVERA, 2008). Moreover, a same class of stochastic models can be used both for synthetic generation and series forecasting, the predicted series, however, not being considered synthetic series.

In addition to deterministic and stochastic models, there is a class of models that adjust the results to the observed data by means of mathematical formulations not related to any physical processes. Those models are classified as empirical models, also called “black-box” models as the relations between input and output are not understood but by mathematical means.

## 1.2 SYNTHETIC HYDROLOGY

Good quality streamflow time series are of paramount importance for successfully fitting a hydrological model. However, the representativeness of such datasets may be a relevant issue to be considered (TAGHI SATTARI et al., 2012).

Time series modelling consists in identifying and reproducing the characteristics of a known series. Moreover, a  $n$  years long synthetic series must keep the statistical parameters of the original regardless of the period of study (RAUDKIVI, 1979). Thus, the first stochastic model for synthetic streamflow generation was proposed by Sudler (1927), in which the streamflow series is understood as a deck of cards and each value corresponds to one card. Synthetic series are then generated by shuffling and randomly reorganizing the cards, assuring the maintenance of the statistical parameters. By contrast, this model ignores the persistence of the series, which is considered to be essential in water resources modelling (DETZEL et al., 2014; DETZEL and MINE, 2016).

Among the many parameters which characterize streamflow series, the most important is the persistence structure, also referred as serial correlation (KELMAN, 1987). High dependence structures are commonly reported for many time series, especially for those related to natural phenomena such as precipitation, wind speed and streamflow. Therefore, the model proposed by Sudler (1927) would no longer be suitable for water resources time series modelling. The persistence of a series is addressed through the autocorrelation function and a strong persistence is mostly inherent to water resources time series. An exception was reported by Guimarães and Santos (2011) at Paiva River, Portugal, that does not present statistically significant temporal dependence at annual scale.

Physically, the persistence characteristic refers to the temporal dependence between the elements of the series, meaning that mainly the flow at the time interval  $t$  is dependent on the flow at the interval  $t - 1$ . Therefore, for a series presenting a high autocorrelation, if in a year  $A_t$  the mean streamflow was above the average, the same behavior is expected in the year  $A_{t+1}$ . The persistence intensity between the element of a series is inversely proportional to the considered temporal scale; thus, the daily streamflow series usually tends to

be more persistent than the monthly as what happens today is more dependent on what happened yesterday than what happened this month is on what happened last month.

Furthermore, the relevance of the persistence for the modelling lead to a wide variety of models that are capable of computing and replicating that structure. Thus, Autoregressive Integrated Moving Average (ARIMA) type models, also known as Box & Jenkins models, were consolidated by Box et.al. (2008) and are still largely used on this purpose by many hydrologists worldwide, presenting a fair representation of the river regime (VALIPOUR et. al., 2013; NEIRA, 2005; PÉRICO, 2014). For monthly time series, where seasons must be considered, there are some ARIMA variations, like PARMA and SARIMA, for which some of the parameters represent seasonality (BAYER et. al., 2012; SHAO et. al., 2009; HALTINER & SALAS, 1988).

The serial correlation inherent to hydrologic temporal series lead the models for synthetic streamflow generation to address the proper representation of such characteristics. Based on that premise, Thomas and Fiering (1962) proposed the known first-order univariate Markov chain – AR(1), expanded by Matalas (1967) to the multivariate case. The AR(1) model is defined by the equation (1):

$$\mathbf{z}_t = \mathbf{A} \cdot \mathbf{z}_{t-1} + \mathbf{B} \cdot \mathbf{a}_t \quad (1)$$

In which  $\mathbf{z}_t = [z_{t,1}, z_{t,2}, \dots, z_{t,l}]^T$  is an array of  $l$  streamflow series, each corresponding to a locality  $u$  ( $u = 1, 2, \dots, l$ ) and  $\mathbf{a}_t = [a_{t,1}, a_{t,2}, \dots, a_{t,l}]^T$  is an array of  $l$  residuals, also corresponding to the localities  $u$ ;  $t$  ( $t = 1, 2, \dots, n$ ) is the temporal index;  $\mathbf{A}$  and  $\mathbf{B}$  are  $l \times l$  parameter matrices.

On the AR(1) model, the matrix  $\mathbf{A}$  is responsible for the modelling of the temporal dependence between the streamflow values, and its estimation is evaluated based on the autocorrelation coefficient of the historical series. Whereas the matrix  $\mathbf{B}$  is responsible for the spatial dependency between the localities, as in order to maintain the coherence of the hydrologic regime of a catchment, the computation of the cross-correlation between streamflows of various locations is, for the multivariate case, of paramount importance.

The AR(1) process for synthetic streamflow generation can be described by three steps: (i) estimate the parameter-matrices **A** and **B** from the historical series; (ii) reverse the equation (1) in order to obtain the residual series  $\mathbf{a}_t$  and apply the theoretical validation of the model over this series; (iii) generate pseudorandom numbers to generate synthetic series  $\mathbf{z}_t$  by using the equation (1). Herein it is important to understand that the aforementioned validation is for the estimated model. It is only after the step (iii) that the validation of the synthetic series is performed through comparison between the descriptive statistics of the historical series and synthetic series. For the AR(1) model, the series must maintain the first and second order statistical moments (mean and variance), the first order serial correlation and the zero-order spatial correlation.

Some aspects regarding the synthetic series generation are of such importance that require a more detailed description. The first feature refers to the pseudorandom numbers, considered responsible for generating the various scenarios. In essence, the algorithms that generate these numbers start from an initial value named seed. Each routine, while using the same seed, will always produce the same sequence of pseudorandom numbers regardless of how many times one runs it. However, diverse sequences are necessary, since the main purpose of synthetic series is to acknowledge the uncertainty (e.g. variability) regarding the historical series. Aiming to address that issue dynamic techniques use a different seed for each synthetic series generation. Computational softwares such as MATLAB have built-in pseudorandom generation functions with the option for also randomizing the seed for a variety of probability distributions. Various methods for the generation of pseudorandom numbers are detailed by Kaviski (2006, appendix A and B).

The second aspect is about the quantity of generated synthetic series, as there is no definitive method to establish the appropriate number. Kelman (1987) state that the quantity must be such that allow the empirical distribution of the synthetic series to be approximately equal to the theoretical distribution of the historical series. The verification, however, is only possible by probabilistic means as the theoretical distribution is unknown. Alternatively, Guimarães and Santos (2011) present a specific analysis on that matter. Their approach regards the synthetic generation of multiple ensembles of streamflow downstream from a reservoir with different quantities of scenarios each. By evaluating the standard



deviation of the storage of the reservoir the authors conclude that that variable is well represented by 1200 synthetic series. Thereafter, Detzel et. al. (2016) conducted a similar research for some hydro plants of the Brazilian National Interconnected System (SIN) considering the maximal accumulated streamflow deficit downstream from the reservoirs. The results suggest diverse quantities depending on the series. The Foz do Areia plant, in Paraná state, required 2000 synthetic series, whereas some required as much as 6000 series.

Finally, the third aspect regards the non-stationarity inherent to hydrological series. The stationary behavior is associated to the statistical independence of these series in relation to time. In short, the mean, variance and autocorrelations do not change over the time in such series. However, there are registers of alterations in those statistics in several hydrological time series. In Brazil, the non-stationarity is mainly present at the South region and at the southern portions of the Southeast and Midwest regions of the country (DETZEL et. al., 2011), characterized from the 1970's decade on. Moreover, a great number of stochastic linear models, including the AR(1), are stationary and thus require the series to present this same behavior. The issue is addressed by the following steps: (i) Evaluate if either a series is stationary or non-stationary; (ii) If non-stationary the trend must be removed. At step (i) hypothesis-testing for stationarity are performed (e.g. linear regression (BORMANN et. al., 2011), t-Student (FILL, 2011), Wilcoxon (THOMAS, 2007), Pettitt (ROUGÉ et. al., 2013), Spearman (FLEMING and WEBER, 2012) and Mann-Kendall (LIANG et. al., 2011)). At step (ii) some classic methods are the graphical method (BATISTA et. al., 2009) and the differentiation of the series (BOX et. al., 2008).

### 1.3 CONSIDERED MODELS

ARIMA type models are parametric as they estimate parameters by statistical analysis (RAUDKIVI, 1979) and those models are particularly good for discharge modelling as they properly reproduce the time dependency intrinsic to streamflow series (DETZEL, 2015). However, a problem regarding ARIMA modelling is that those classic models are linear and fail to represent the non-linearity of streamflow and, thus, might not give the best result. An alternative approach, possible to improve the quality of results and adequate to deal with non-linear patterns is the artificial neural network (ANN) (AHMED AND SARMA,

2007). ANN models are trained to represent the processes and relationships characteristic of each data set, computing information similarly to the biological nervous system with a high number of “neurons” working in parallel. Many studies have been taken with ANN hybrid models in order to better compute both linear and nonlinear trends (FARUK, 2010; CARNEIRO & FARIAS, 2013; ABRAHART & SEE, 2000).

### 1.3.1 Box & Jenkins Models

The term Box & Jenkins Models refers to ARIMA models consolidated by Box et.al. (2008). This group of models is obtained from a linear combination of an autoregressive portion (AR), an integration factor (I) and a moving average portion (MA). Therefore, it is possible to model a series by using any combination of these three elements. The non-seasonal model is denoted as ARIMA (p,d,q), where p, d, q are non-negative integers representing respectively the orders of autoregressive, integration and moving average portions.

The formulation of this ensemble of models allow them to properly compute and replicate the persistence of a series, that being the main reason why they are still largely used on this purpose by many hydrologists worldwide (VALIPOUR et. al., 2013; NEIRA, 2005; PÉRICO, 2014). When formulating an ARIMA (p,d,q) model, a very useful notational device is the backward shift operator B, defined by the equation (2).

$$z_{t-k} = B^k \cdot z_t \quad (2)$$

The ARIMA models, defined by the equation (3), have shown to accordingly approach the serial correlation and efficiently address the descriptive statistics of the historical series, therefore presenting a fair representation of the river regime. However, in order to correctly use this group of models it is crucial to take into account the differences between modelling series with stationary and non-stationary behaviors. The stationary models suppress the integration factor, being a combination of AR and MA portions, whilst the integration factor enables the model to deal with the non-stationarity.

$$\phi(B) \cdot (1 - B)^d z_t = \theta(B) \cdot a_t \quad (3)$$

$$\Phi(B) = 1 - \sum_{i=1}^p \phi_i B^i$$

$$\Theta(B) = 1 - \sum_{i=1}^p \theta_i B^i$$

in which  $B$  is the backshift operator,  $\phi_i$  and  $\theta_i$  are the  $i$  order parameters of respectively the autoregressive and the moving average models, and  $a_t \sim N(0,1)$ . The observed series may be submitted to logarithmic or other numerical transformation in order to comply with the normality requirement. Such condition is a premise of the ARIMA model.

#### 1.3.1.1 Autoregressive Models (AR)

The autoregressive model (AR( $p$ )) is the ARIMA model with the Integration and Moving Average portions of order zero. Hence, it can also be denoted as ARIMA( $p, 0, 0$ ). Hydrologists have been using this type of model since the 1960's, (e.g. Matalas, 1967) as it fairly represents events at annual and smaller scales, nevertheless the modelling process is more complex for small scales as seasonality must be considered (SALAS et. al., 1980). With  $d$  and  $q$  equal to zero the equation (3) is rewritten as equation (4):

$$\Phi(B) \cdot z_t = a_t \quad (4)$$

Autoregressive models are still widely used for annual streamflow generation, as in the study of Neira (2005) where the synthetic series were used for risk reduction in reservoir operation. Périco (2014) studies the influence of reservoirs in hydroelectric power generation using synthetic series generated by AR models.

#### 1.3.1.2 Autoregressive Moving Average models (ARMA)

In this formulation, a Moving Average portion is added to the AR, in a model called ARMA ( $p, q$ ), corresponding to an ARIMA ( $p, 0, q$ ). The Moving Average portion (MA( $q$ )) describes the relation between the model residuals, as shown in the equation (5).

$$\Theta(B) \cdot a_t = z_t \quad (5)$$

The combination of the equations (4) and (5) is defined by the equation (6) and characterizes an ARMA(p, q) model.

$$\phi(B) \cdot z_t = \theta(B) \cdot a_t \quad (6)$$

These stationary models have several significant theoretical properties regarding the variance and the autocorrelation. Two among these, namely the stationarity and invertibility of the model, are detailed herein for they are essential for configuring the ARMA model. Furthermore, the chapter 3 of Box et. Al. (2008) addresses those properties.

Considering an AR(p) model, for the stationarity condition to be established, the roots of the characteristic polynomial  $x^p - \phi_1 \cdot x^{p-1} - \phi_2 \cdot x^{p-2} - \dots - \phi_p = 0$  must lie inside of the unit circle. For the AR(1) model, the polynomial results in  $x - \phi_1 = 0$ , therefore, for the root to lie inside of unit circle  $|\phi_1| < 1$ . This verification for the stationarity condition is unnecessary for MA(q) models.

The invertibility of a model refers to the capability of rewriting a MA process in order to obtain a pure AR process as illustrated hereon by using a first order moving average model (BOX et. Al., 2008, p. 52):

$$z_t = (1 - \theta B) \cdot a_t \quad (7)$$

Expressing  $a_t$  in terms of present and past  $z_t$ 's, the equation (7) becomes  $a_t = (1 - \theta B)^{-1} \cdot z_t$ , therefore:

$$a_t = (1 + \theta B + \theta^2 B^2 + \dots + \theta^k B^k)(1 - \theta^{k+1} B^{k+1})^{-1} z_t \quad (8)$$

$$z_t = -\theta z_{t-1} - \theta^2 z_{t-2} - \dots - \theta^k z_{t-k} + a_t - \theta^{k+1} z_{t-k-1}$$

On letting  $k \rightarrow \infty$  one obtain:

$$z_t = -\theta z_{t-1} - \theta^2 z_{t-2} - \dots - \theta^k z_{t-k} + a_t \quad (9)$$

For equation (9) to converge,  $|\theta| < 1$ ; in that case the process is considered to be invertible. Thus, analogously to the stationarity, for a MA model to be invertible the roots of the characteristic polynomial  $\theta(B) = 0$  must lie inside of the unit circle and this verification for the invertibility condition is unnecessary for AR(p) models.

In conclusion, when considering an ARMA model, the restrictions for conditions of both stationarity and invertibility must be fulfilled. Consequently, the roots of both  $\phi(B)$  and  $\theta(B)$  must lie inside of the unit circle.

#### 1.3.1.3 Autoregressive Integrated Moving Average models (ARIMA)

ARIMA  $(p, d, q)$  models are the general form of ARMA  $(p, q)$ . Those models allow the computing of homogeneous non-stationary series by adding an integration factor (I) to the formulation (Hipel e McLeod, 1994, p. 76). The model is defined by the equation (10):

$$\phi(B) \cdot w_t = \theta(B) \cdot a_t \quad (10)$$

in which  $w_t = \nabla^d z_t$ . The operator  $\nabla^d$  can be expressed in terms of the backshift operator  $B$  by considering  $\nabla^d = (1 - B)^d$ . Thus, the equation (10) is rewritten in terms of  $z_t$  as the equation (11) and presents the general formulation of the ARIMA model:

$$\phi(B) \cdot (1 - B)^d \cdot z_t = \theta(B) \cdot a_t \quad (11)$$

The conditions for stationarity and invertibility for the ARIMA model are the same as those for the ARMA model but for a slight change for the stationarity. For non-stationary homogenous series, for an ARIMA $(p, d, q)$ ,  $d \geq 1$  roots of the characteristic polynomial  $\theta(B) = 0$  must lie on the unit circle, whilst  $p - d$  roots lie inside of it.

#### 1.3.1.4 The iterative approach to model building of Box & Jenkins

The procedure to correct building the ARIMA models to a hydrologic series passes through three stages, namely: (i) Identification; (ii) Estimation; (iii) Validation (BOX et. Al., 2008). The iterative approach to model building for forecasting and control proposed by these authors is illustrated by Figure 3. This same approach may also be adopted for synthetic series generation.

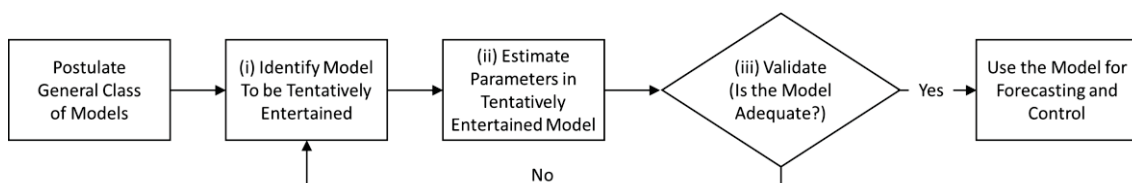


Figure 3. Stages in the iterative approach to model building. (Font: Box et. Al., 2008)

Primarily, the most common metric for the identification of the ARIMA model to be used is built on the graphic comparison between the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) plots. A lag  $k$  autocorrelation is given by the ratio of covariance ( $\hat{\gamma}_k$ ) to sample variance ( $\hat{\sigma}_z^2$ ), as given by the equation (12). The chart of the coefficient  $\hat{\rho}_k$  versus the lag  $k$  is known as the ACF and illustrates the relation between the elements of a series.

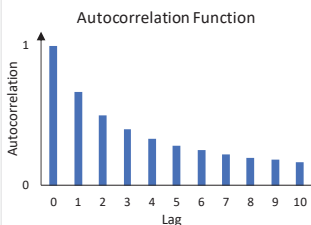
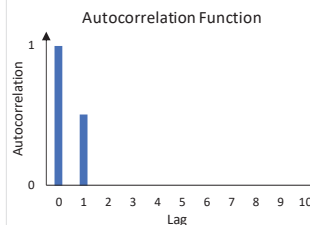
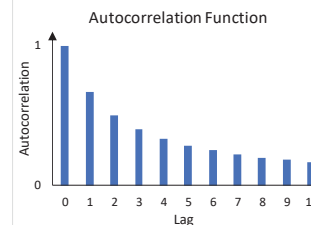
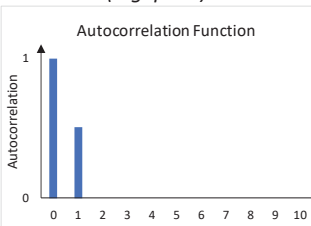
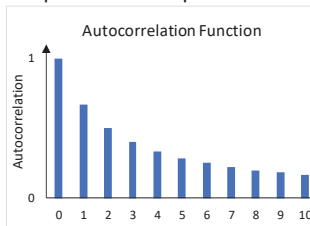
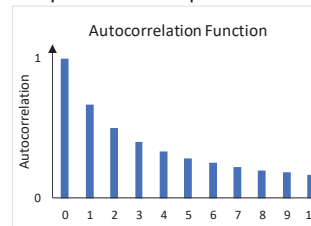
$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\sigma}_z^2} = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} \quad (12)$$

The PACF, noted as  $\varphi_{kk}$ , is a complementary function for the understanding of the dependence between the series elements and is given by the equation (13). The calculation is recursive and the set of parameters ( $\varphi_{11}, \varphi_{22}, \dots, \varphi_{kk}$ ) characterizes the partial autocorrelations.

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\sigma}_z^2} = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} \quad (13)$$

The ACF and PACF are evaluated and plotted and their charts are compared with the theoretical expected behaviors for AR(p), MA(q) and ARMA(p,q) models (HIPEL and MCLEOD, 1994; BOX et. Al., 2008). Those expected behaviors were plotted by Souza and Camargo (2004) and are illustrated at Table 1. One may notice that for the AR portion the ACF behavior specifies whether this parcel is relevant to the model, whereas the PACF designates the order of this portion. For the MA the opposite is true, with the PACF behavior related to the applicability of the parcel, whilst the ACF indicates the order. In essence, in cases when autoregressive patterns are pertinent for the modelling, the ACF presents characteristics close to an exponential decay. By contrast, the same behavior observed in the PACF denotes a meaningful moving average process in the series. The graphical analysis method specifies that the order of each model (AR and MA) is determined by the number of lags of significant autocorrelation in the PACF for the prior and the ACF for the later.

Table 1. ACF and PACF theoretical behavior

Type of Model	AR ( $p$ )	MA ( $q$ )	ARMA ( $p, q$ )
Typical Pattern of ACF	Decays exponentially or with damped sine wave pattern or both 	Cut-of after lag $q$ (e.g. $q = 1$ ) 	Decays exponentially or with damped sine wave pattern or both 
Typical Pattern of PACF	Cut-of after lag $p$ (e.g. $p = 1$ ) 	Decays exponentially or with damped sine wave pattern or both 	Decays exponentially or with damped sine wave pattern or both 

Font: Adapted from Souza and Camargo (2004)

When trying to identify the model to be entertained, the graphic methods present a quick analysis but, although they present decent results, these methods are subjective and should be complementary only. Therefore, a mathematical procedure is required for a computer to be capable of precisely identify a proper model. One feasible metric for model identification is the Bayesian Information Criterion (BIC) (SCHWARTZ, 1978), given by the equation (14):

$$\text{BIC} = -2 \cdot \ln L(\hat{\phi}, \hat{\theta}, \hat{\sigma}_a | z_t) + \ln n \cdot (p + q) \quad (14)$$

where  $\ln L(\hat{\phi}, \hat{\theta}, \hat{\sigma}_a | z_t)$  is the log-likelihood function (to be presented further in this work),  $n$  is the length of the time series  $z_t$  and  $p$  and  $q$  are respectively the autoregressive and moving average orders for the model. The method consists of evaluating equation (14) for all the postulated models and select the one resulting the lowest BIC. For more details on the BIC procedure, one may refer to Schwartz (1978).

Once the model to be used is chosen, one must estimate the parameters  $\phi_i$  ( $i = 1, 2, \dots, p$ ) and  $\theta_j$  ( $j = 1, 2, \dots, q$ ) respecting the invertibility and stationarity conditions (Box et. al., 2008). Thus, all the roots of the polynomials presented in equations (15) and (16) shall remain inside the unit circle in order to uphold the invertibility and stationarity.

$$x^p - \phi_1 x^{p-1} - \dots - \phi_p x^0 = 0 \quad (15)$$

$$x^q - \phi_1 x^{q-1} - \dots - \phi_q x^0 = 0 \quad (16)$$

There are many methods to estimate the parameters, being the solution of Yule-Walker equations (i.e., equation (17)) for  $\rho_p \cong \hat{\rho}_p$  considered efficient for AR (p) models.

$$\begin{aligned} \rho_1 &= \phi_1 + \phi_2 \rho_1 + \dots + \phi_p \rho_{p-1} \\ \rho_2 &= \phi_1 \rho_1 + \phi_2 + \dots + \phi_p \rho_{p-2} \\ &\vdots \\ \rho_p &= \phi_1 \rho_{p-1} + \phi_2 \rho_{p-2} + \dots + \phi_p \end{aligned} \quad (17)$$

For ARIMA (p,d,q) models, Box et. al. (2008) suggest the use of the maximum likelihood estimates as an efficient method. This method targets the optimal set of parameters to associate the model results with the observed values. Assuming the hypothesis of normally distributed data, the likelihood function for a time series  $z_t (t = 1, 2, \dots, n)$  and parameters  $\phi$  and  $\theta$  would be:

$$L(\hat{\phi}, \hat{\theta}, \hat{\sigma}_a | z_t) = \frac{1}{(2\pi\hat{\sigma}_a)^{n/2}} \cdot \exp \left[ -\frac{1}{2\hat{\sigma}_a} \cdot \sum_{t=1}^n \hat{a}_t(\hat{\phi}, \hat{\theta})^2 \right] \quad (18)$$

where  $a_t$  is the residual series with a sample standard deviation  $\hat{\sigma}_a$ . As a matter of simplification on the mathematical process the log-likelihood equation associated with the equation (18) is given by:

$$\ln L(\hat{\phi}, \hat{\theta}, \hat{\sigma}_a | z_t) = -n \cdot \ln \hat{\sigma}_a - \frac{SSQ(\hat{\phi}, \hat{\theta})}{2\hat{\sigma}_a} \quad (19)$$

where  $SSQ(\hat{\phi}, \hat{\theta}) = \sum_{i=1}^n a_i^2$  is the sum of square function for the residuals. Thus, a set of parameters to maximize equation (19) or minimize the  $SSQ(\hat{\phi}, \hat{\theta})$  gives the maximum likelihood for the model. The solution involves an iterative process and must respect the models stationarity and invertibility conditions.

The model validation checks the residuals for temporal independence, homoscedasticity and normality. For the temporal independence the Portmanteau type test (LI and MCLEOD, 1981) is a suitable option. The statistics



of the test is defined by equation (20) and the null hypothesis is rejected if  $Q_L > \chi^2_{\alpha, (L-p-q)}$  for a significance of  $\alpha$ .

$$Q_L = n \cdot \sum_{k=1}^K \rho_k^2(\hat{a}) + \frac{K(K+1)}{2n} \quad (20)$$

in which  $\rho_k$  is the ACF of the series residuals,  $K$  is the maximum lag for the ACF evaluation, between 15 and 25 and not exceeding  $n/4$ . The homoscedasticity can be tested through the Levene test (BROWN and FORSYTHE, 1974) performed on multiple samples ( $g$  samples). The null hypothesis is  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_g^2$  and the statistics of the test is defined by equation (21) rejecting the null hypothesis for  $W_0 > F_{\alpha, g-1, n-g}$ .

$$W_0 = \frac{\sum_{i=1}^g n_i (a_i - \bar{a})^2 / (g-1)}{\sum_{i=1}^g \sum_{j=1}^{n_i} (a_{ij} - \bar{a}_i)^2 / \sum_{i=1}^g (n_i - 1)} \quad (21)$$

Lastly, the residual series  $a_t$  must be normally distributed and several tests (e.g., Chi-Squared; Kolmogorov-Smirnov; Shapiro-Wilk; Jarque-Bera) are suitable for this verification. If any of the premises is not satisfied, one may select a new model and proceed with the iteration process from the beginning.

### 1.3.2 Artificial Neural Network Models

Neural networks are a branch of artificial intelligence (AI) that similarly to the human brain process information through non-linear parallel synapses, which, in essence, consists of subdividing a complex problem into a group of relatively simpler tasks. Some highlights of ANN are the nonlinearity, the self-learning capability, the adaptability and the response to evidences (HAYKIN, 1999).

The AI has a wide range of applications in engineering problem solving, and has increased since the 1980s (PRADA-SARMIENTO & OBREGÓN-NEIRA, 2009). Thus, ANNs has been used for more than 20 years in water resources related fields, becoming a deep-rooted research area and showing a substantial progress in the last two decades in forecasting and modelling non-linear parameters with a fair representation of the noise complexity (MAIER et al., 2010; ZHANG et al. 2018; YASEEN et al., 2015). ANNs tend to be particularly useful

when applied to complex processes, where the details of which are not well understood (SCHMID et al., 2006).

The ANN mechanism relies upon a set of processing units called neurons disposed in arrays called layers and interconnected by synapses. The synapses are links of variable weights between neurons of different layers. The feed-forward architecture is the most common among Neural Network (NN) based models (AHMED and SARMA, 2007). On those architectures the synapses occur in one direction on a layer-by-layer basis. Moreover, feed forward networks have one input layer which receive the data, also called the stimulus of a NN, one or more hidden layers where the data are processed and one output layer, responsible for the response of the NN. This architecture scheme is illustrated in Figure 4 for a two hidden layer structure.

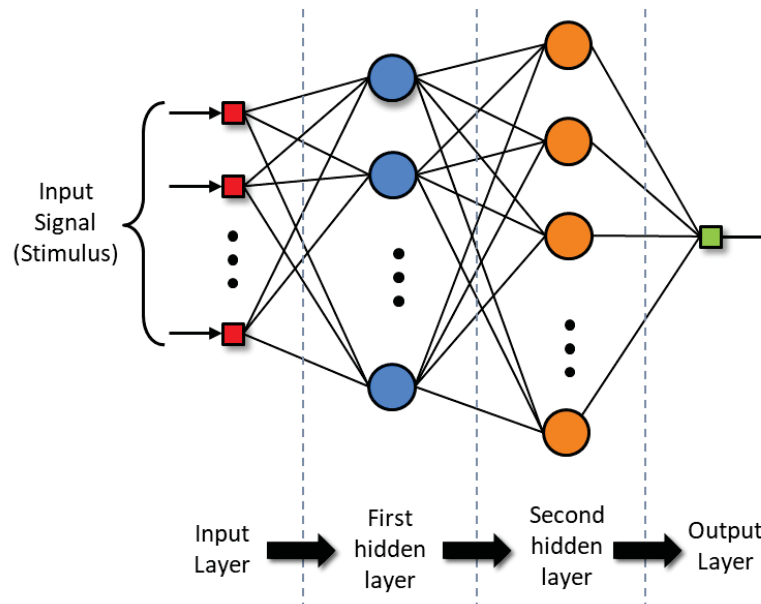


Figure 4. General architecture of a two hidden layers ANN

A right architecture is crucial for obtaining a satisfactory ANN based model. The choice of relevant parameters for input and the right number of neurons in each layer are of paramount importance and can be determined by a trial-and-error process or by using optimization algorithms (AKSOY and DAHAMSHEH, 2018; GALVÃO et al., 1999; HAYKIN, 1999). According to Maier et al. (2010) Multilayer Perceptrons (MLPs) are the most common form of feed-forward model architecture, whereas the back propagation is the most commonly used supervised training algorithm in multilayer feed forward networks (ZHANG et al., 2018; ZADEH et al. 2010). Other feed-forward network architectures in use include Generalized Regression Neural Networks (GRNNs), Radial Basis

Function (RBF) networks, Neuro-fuzzy networks and Support Vector Machines (SVMs). This dissertation addresses the MLPs only.

The process of training a NN aims to calibrate the network based on the given data. Fundamentally, the Neural Network processes the data through synaptic weights and biases. By working with inputs paired with expected outputs, the calibration consists in estimating such weight values capable to simulate outputs statistically similar to the expected. Specific algorithms named back-propagation algorithms optimize the set of weights and biases reducing the deviation between expected and simulated outputs. Additionally, some pre-determined factors directly affect the quality of the training. The ANN architecture, the activation function, the number of iterations for the training algorithm to be performed, the portion of the original series used for training and the initialization weights have a strong influence on the results, its choice, however, consists in a challenging task. Herein, care should be taken during the training in order to ensure its adequacy while avoiding over complexity (ADELOYE, 2009; FERREIRA et al., 2011; HAGHIABI, 2017; MACHADO et al., 2011; PRADA-SARMIENTO and OBREGÓN-NEIRA, 2009).

The architecture of a Multilayer Perceptron Neural Network (MLP) is characterized by one input layer, one or more hidden layers and one output layer. Each layer is composed of one or more neurons and the neurons are connected to the following layer through weighted synapses, with the propagation of the input vector occurring on a layer-by-layer basis in a forward direction. In essence, a three layer (i.e. input, output and one hidden layer) architecture is suitable for any non-linear function (GALVÃO et al., 1999).

According to Haykin (1999), there are three characteristics of MLP that are evident in a network, namely:

- (i) Each neuron of each layer has a nonlinear activation function, being the sigmoidal nonlinearity presented on equation (22) a commonly used one:

$$\varphi_j = \frac{1}{1 + \exp(-v_j)} \quad (22)$$

in which  $v_j$  is the induced local field (i.e. the weighted sum of all synaptic inputs plus the bias) of the neuron  $j$ , and  $\varphi_j$  is the output of the neuron  $j$ .

- (ii) There are one or more layers other than the input and the output layers. Those neurons are responsible for learning from the input and progressively improving results through the training.
- (iii) The synapses of the network exhibit elevated degree of connection between the neurons, which means that even small changes in the architecture of the network might require a change in the synaptic weights.

An example of a Multilayer Perceptron with a one hidden layer architecture is given in Figure 5.

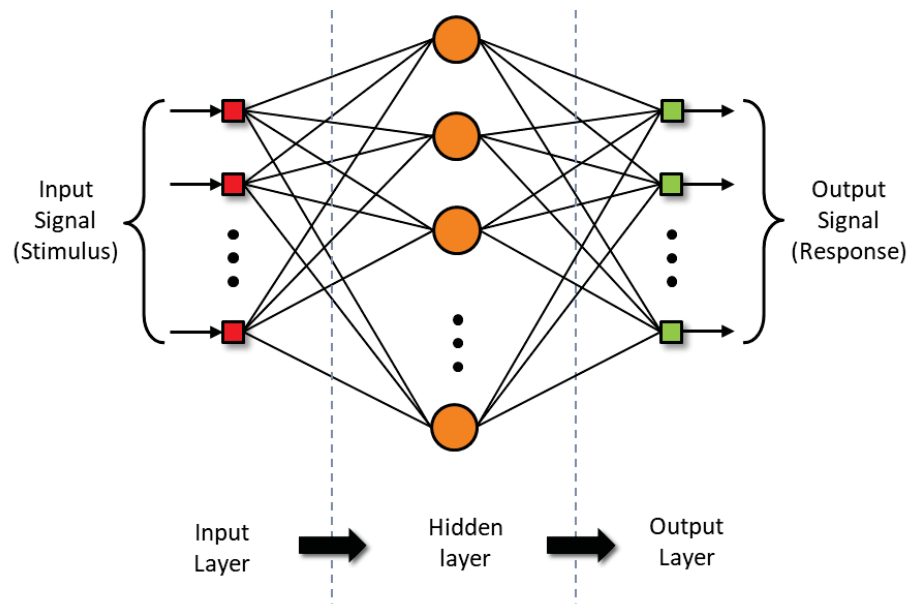


Figure 5. Architecture of a Multilayer Perceptron with one hidden layer

Neurons from subsequent layers are fully connected and the signal flows are unidirectional, progressing on a layer-by-layer basis. Thus, there are two types of signals that propagate through an MLP, the function signals and the error signals. Function signals, also referred as Input Signals, are originated at the input layer and spread forward through the layers of the network, resulting in an output signal at the end of the network. Conversely, error signals propagate backwards through the network, originating at the output layer and involving error-dependent functions for its computation by each neuron.

The architecture of a MLP characterizes a nonlinear model which can suit for modelling nonlinear variables such as the hydrological variables, some

authors also highlight the MLP as a traditional, commonly used and one of the most popular neural networks among hydrologists (COULIBALY et al., 2001; SHOAIB et al., 2016; ZHANG et al., 2018). Therefore, several water resources related studies were taken using those neural networks in order to evaluate a wide range of hydrological variables in a variety of approaches. Furthermore, this architecture, although largely used on hydrology, frequently presents sub-optimal results (COULIBALY et al., 2001).

Some authors use MLP models for the rainfall-runoff transformation. Machado et al. (2011) model and compare an ANN-based empirical rainfall-runoff model with a deterministic model. The modelling regards the Jangada River basin in the state of Paraná, Brazil. The processes involved in the transformation of rainfall in runoff make it of high complexity, however the neural networks showed to be efficient in detecting those patterns and performed better than the conceptual model. The authors conclude that the length of the data series is directly proportional to the number of inputs, even though they acknowledge that this should be further investigated.

Zadeh et al. (2010) model a Multilayer Perceptron Neural Network which has Rainfall and Runoff data as inputs in order to forecast daily flows in a humid tropical river basin with a strong seasonal rainfall pattern, as they affirm that the relationship between those two variables is essential on flood prediction. Moreover, the applicability of two sigmoid activation functions is compared and they infer that, for the study area, the tangent sigmoid activation function (Equation (22)) performs better than the logistic sigmoid activation function (Equation (23)). In conclusion, as the choice of the input vector directly impacts the quality of the result their study suggest that the correlation analysis is sufficient to determine the best fitting input vector to the MLP.

$$\varphi_j = \frac{2}{1 + \exp(-2v_j)} \quad (23)$$

The study of Ochoa-Rivera (2008) improved the ANN model with an stochastic approach, in the sense that the author proposes a non-linear multivariate MLP based model with a normally distributed random factor in order to obtain synthetic annual drought scenarios. The method consists of a three-layer network that uses the monthly runoff data from multiple catchments of the

same basin as input and generates monthly streamflow as outputs, used then to obtain synthetic annual droughts. Furthermore, the model is applied in a basin located in central Spain, using data of seven streamflow stations, thereafter the results were compared with those obtained by a second order autoregressive model – AR(2). The results show that the performance of the non-linear model is significantly better than the linear approach, which the author attributes to the non-linearity of the first.

Schmid et al. (2006) lead a study on dissolved oxygen on free water surface ponds using an MLP based model applied to a wetland pond in southern Finland. The study investigates the performance and advantages of using ANNs on modelling a process which the author specifically describes as a complex function of several hydrological, hydrodynamic and ecological variables. For the study, hourly data of the water temperature vertical profiles, turbidity and oxygen saturation, inflow and outflow rates, as well as wind speed and direction were collected. The study showed the wind effect to be negligible for this study, thus, the remaining parameters, in addition to the hour of the day, were used as inputs of a three-layer network, with the 4-hour forecast of oxygen saturation being the only output of the model. Similarly, a study about the longitudinal dispersion coefficient prediction in natural streams was made by Haghiabi (2017) using MLP in the water quality field. The study compares a multivariate adaptive regression splines (MARS) method with an MLP.

Sudheer (2005) states that important information on the physical processes characteristic of a data set are rooted inside black-box models, on top of that, Prada-Sarmiento and Obregón-Neira (2009) study the mathematical relationship between the synaptic weights of an MLP and some geomorphological features of watersheds applied to the central region of Colombia. The author trained the model for some different catchments, obtaining their respective weights. Afterwards, the author uses some statistical inference techniques to relate each matrix of weights to characteristics such as area, slope and length of the main river of the respective basin.

## 2 HYBRID MODELS AND PROPOSED SCHEME

When searching for Artificial Neural Network on the hydrology top ranked scientific periodic, one can notice that a high frequency of Hybrid models combining ANN with some conventional statistical method for a wide range of purposes on the Hydrology field. It is usual to deal with time series containing both linear and nonlinear patterns. Therefore, it is expected that neither a linear model nor a nonlinear neural network alone can optimally compute both patterns. Recent studies have shown that the accuracy of hybrid linear-nonlinear models results is improved in comparison with those obtained by single models (ÖMER FARUK, 2010; YASEEN et al., 2015).

Nguyen-ky et al. (2017) associate an ANN with a Bayesian approach aiming the water market pricing in Australia's Murry Irrigation Area. The hybrid model has shown to be capable of computing complex non-linear processes and this model has performed better than the single ANN model. Alternatively, Ömer Faruk (2010) obtained significantly better results for a hybrid ARIMA-ANN model over single ARIMA and single ANN models for water quality predictions at Büyük Menderes River.

Khashei and Bijari (2010) construct a hybrid ARIMA-ANN model aiming to improve the forecasting for nonlinear series, which understands each value as a nonlinear function of past events. Therefore, the model primarily uses an ARIMA model in order to generate synthetic data, which will subsequently be used by a neural network to predict future values of time data. According to the authors, coupling divergent models in a hybrid model may reduce the model error and uncertainty as well as improve its performance. This statement corroborates with the results in which the proposed ARIMA-ANN model overcomes the non-hybrid models.

The methods diverge from one author to another, but all of them convey the idea that hybrid models can perform better than single models due to presenting a more comprehensive approach. This indicates that hybrid models might be a good choice for modelling the complexity of hydrological time series. Thus, this research proposes coupling the well-established ARIMA model with an ANN based model for synthetic streamflow generation. The choice for the ARIMA model mainly relays on its capability of addressing the persistence of a series.

Moreover, besides its linear formulation, the Box and Jenkins set of models has shown a good performance in water resources modelling over the past decades. Due to numerical issues regarding the integration factor of the model for synthetic generation, the focuses of the ARIMA portion of the hybrid model will be the AR and ARMA models. For the ANN portion, a MLP is chosen and is intent to be trained in such way that fairly compute the non-linearity and the trend of the historical series.

There are many studies that use hybrid ARIMA and ANN models. However, there is no consensus on how to couple the models and each study hybridizes both models by its own methods. That being said, a new method of coupling was thought specifically for this research, as follows: Under the hypothesis that passing a non-stationary series through the linear formulation of an ARMA filter not only the non-linearity but also the non-stationarity of the series will be kept within the residual series.

Preserving the non-linearity and the non-stationarity for the residual series and using the residual series as an input of the ANN, the network should be able to detect both non-linear and trend patterns of the series. Aiming at that, a preliminary analysis was made and is detailed at the chapter 4.1, in which the non-stationarity was confirmed to be preserved for the residuals. Thus, the developed scheme for coupling the ANN to the ARMA model for a series  $z_t$  can be simplified by three steps: (i) The ARMA model receives the historical series  $z_t$  in order to estimate the parameters of the model and produce the residual series  $a_t$ ; (ii) The residuals are used as input for the implementation and training of the ANN, which thereafter should be capable of generating a group of synthetic residual series  $a_t^*$ ; (iii) The set of synthetic residual series return to the ARMA model to generate synthetic streamflow series  $z_t^*$ . A simplification to this scheme is presented by Figure 6 and the process is thereafter described in detail.



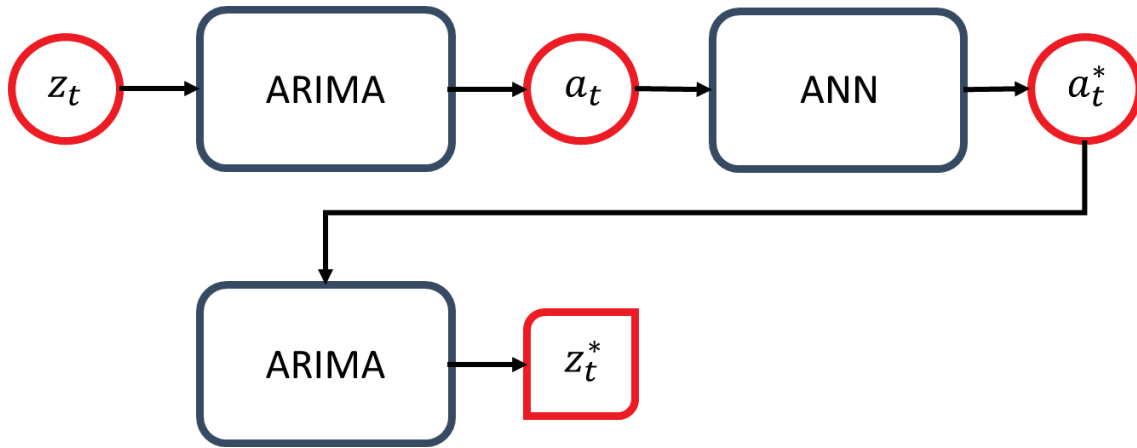


Figure 6. ARMA-ANN coupling scheme

Considering the monthly scale, the hydrological time series  $z_t$  must be deseasonalized, for the ARMA formulation assumes as a non-seasonal model. Subsequently, the iterative approach to model building of Box & Jenkins, as described in the chapter 1.3.1.4, is performed in order to identify, estimate the parameters of and validate the model. The study was taken for six stations and the identification was made by means of the BIC, in order to reduce the subjectivity and automatize the results. Herein, it is important to estipulate maximum orders for AR and MA portions of the model. Therefore, the postulated models should have orders for the AR and MA parts ranging from zero to 2, for keeping it parsimonious and the integration portion would be suppressed as it is suitable for forecasting rather than synthetic generation.

Having identified the most suitable among the postulated models, the following step estimates the parameters by means of the maximum likelihood estimates method. It was chosen because of its comprehensive approach and considerable efficient results. Subsequently, considering the orders  $p \leq 2$ ,  $q \leq 2$  and  $d = 0$ , the equation (11) is rewritten as equation (24), in order to obtain the residual series. Before using the residual for implementing the neural network, the independence, homoscedasticity and normal distribution should be checked and the Portmanteau, Levene and Jarque-Bera tests were performed.

$$a_t = z_t - \phi_1 \cdot z_{t-1} - \phi_2 \cdot z_{t-2} + \theta_1 \cdot z_{t-1} + \theta_2 \cdot z_{t-2} \quad (24)$$

Afterwards, the residual series is used for the training and validation of the Neural Network. The considered type of architecture is a one hidden layer MLP, with a one neuron output layer  $a_{t+1}^*$  representing the residual for the month

$t + 1$ . For the input, the network considers the array of residuals for the previous  $k$  months ( $A = [a_{t-k+1}, a_{t-k+2}, \dots, a_t]$ ) for obtaining the next month “residual”. The most suitable architecture is obtained by a trial and error method, with the  $k$  inputs varying from 1 to 18 months and the neurons  $n$  on the hidden layer ranging from  $i = 1$  to 36, as shown in Figure 7. A portion of 70% of the data is used for the training, while the remainder 30% is considered for the validation. The performance is evaluated by the root mean square error (RMSE) and determination coefficient ( $R^2$ ) metrics. Moreover, the Levenberg-Marquardt algorithm (MORÉ, 1978) is used for the training. Finally, the analysis for the RMSE indicates the architecture to be used in order to generate the synthetic series.

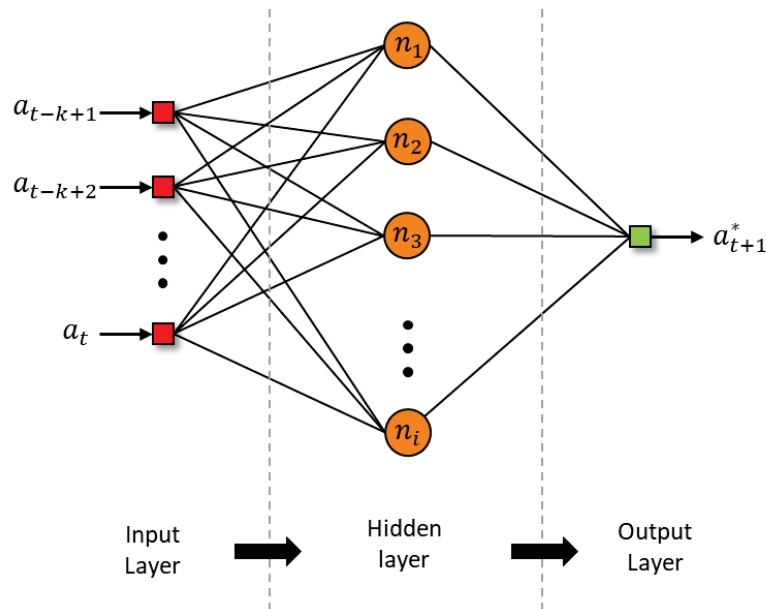


Figure 7. MLP tested architectures

From the most suitable architecture among those tested, a recursive method is used to produce synthetic values. In essence, the last  $k$  values of residuals are used to generate the following residual ( $a_{t+1}^*$ ). At each iteration  $t \leftarrow t + 1$ , and the last calculated output becomes one input for the following synthetic residual. The process continues for as long as the synthetic series size ( $S$ ) times the number of series to be generated ( $N$ ), producing one continuous series that returns to the ARMA equation producing a long sequence of synthetic streamflow series. This sequence is thereafter divided in  $N$  synthetic series, those of which are tested and compared with the historical and single ARMA synthetic series by both short- and long-term statistics (DETZEL et al., 2014) means.

Furthermore, the statistical tests of Mann-Kendal (LIANG et. al., 2011) and Pettitt (ROUGÉ et. al., 2013) for stationarity are performed in order to verify the capability of the Neural Network in computing and replicating the trend of the series.

### 3 STUDY AREA

The Iguaçu River basin is located within the Brazilian states of Paraná and Santa Catarina and the Argentine province of Misiones. It is one of the sub-basins of the Paraná River system together with the basins of Paranapanema, Tietê, Grande and Paranaíba as shown in Figure 8.

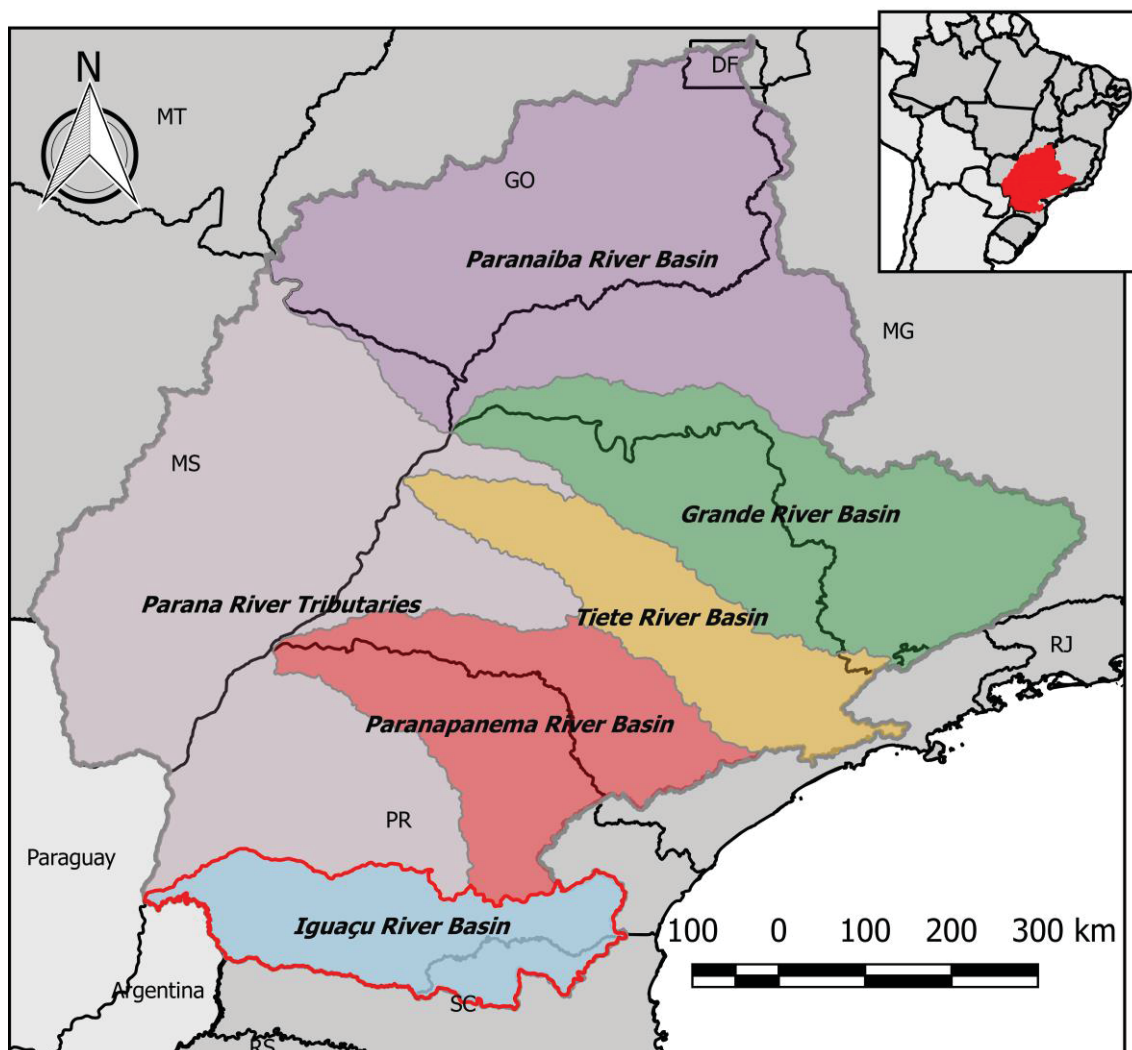


Figure 8. Map of the Parana River basin

This basin is relevant for the Brazilian hydropower exploitation, and thus for this study there were selected six gauging stations located downstream six hydroelectric plants within the Iguaçu river basin. The series are all naturalized, continuous and of 85 years in length.

This chapter addresses the physical characteristics and the climate at the Iguaçu River basin. Additionally, some relevant information about the selected gauging stations are detailed.

### 3.1 CHARACTERISTICS

The Iguaçu river basin is situated between the coordinates 25°05'S and 26°45'S of latitude and the degrees 48°57'W and 54°50'W of longitude, covering about 70.800 km<sup>2</sup> within the Brazilian states of Paraná and Santa Catarina, as showed in Figure 10, and the Argentine province of Misiones. Moreover, the basin is designated by the National Water Agency of Brazil (Agência Nacional de Águas, ANA) by the number 65. Iguaçu River starts on the encounter of Iraí and Atuba rivers, between the municipalities of Curitiba and Pinhais, at an elevation of around 1200 m. It stretches across the state of Paraná, draining to the Paraná River 910 km downstream on the border between the Brazilian city of Foz do Iguaçu and the Argentine city of Puerto Iguazú, at an elevation of around 300 m. This considerable difference between the river head and its falls combined with the area of the basin, enabled the construction of six big hydroelectric power plants within the Iguaçu cascade and many smaller plants in its tributaries.

The Brazilian state of Paraná comprehends 80,4% of the Iguaçu Basin, which represents about 57.000 km<sup>2</sup>. The Iguaçu River crosses the state from east to west and its basin is commonly divided in three sub-basins, namely Upper, Intermediate and Lower Iguaçu (Figure 9). The most important tributaries are Negro, Potinga, da Areia, Iratim, Jordão, Cavernoso, Chopim, Guarani, São Salvador and Capanema.

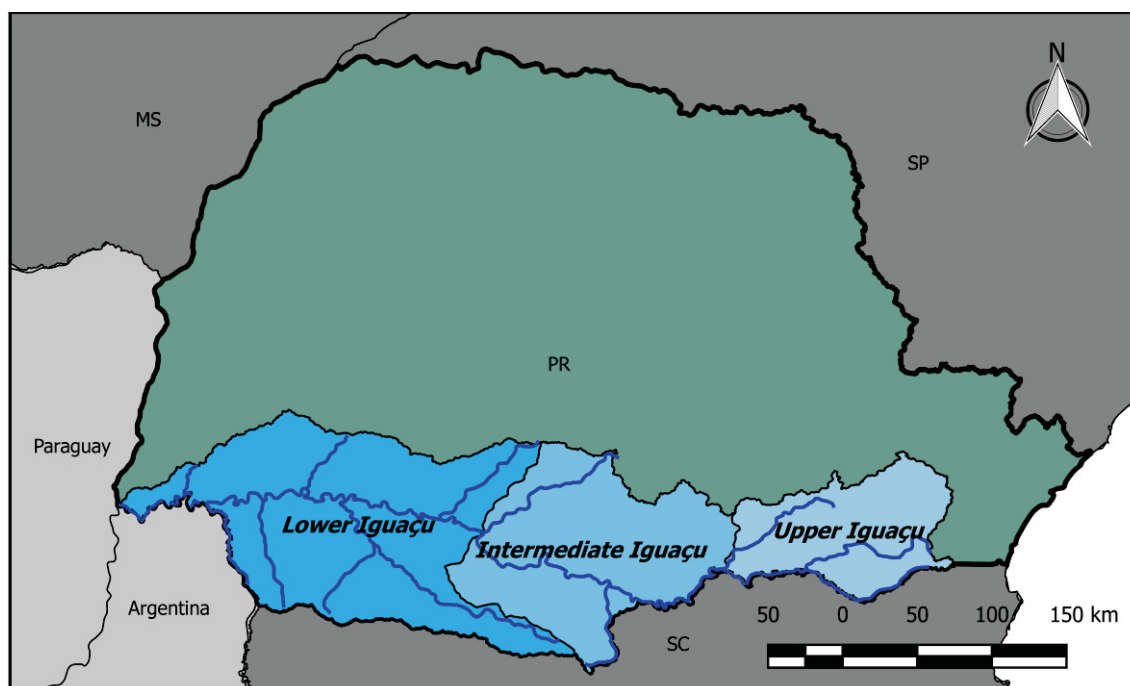


Figure 9. Map of the Iguaçu River basin, divided in Upper, Intermediate and Lower Iguaçu

The Upper Iguaçu, where the Curitiba Metropolitan Area is located, is characterized by intense industrial and commercial activity and elevated population density. Alternatively, the Intermediate and Lower Iguaçu present extensive agricultural activity and are relevant for their hydroelectric potential.

### 3.2 CLIMATE

The Iguaçu basin is inserted in a moist subtropical mid-latitude climate region, with mild to cold winters and warm to hot summers. According to the Köppen-Geiger criteria, the basin is almost entirely classified under the Cfa climate type, but for a small portion at the Upper Iguaçu Basin classified as Cfb, meaning that the basin is in a temperate region, without dry season, with hot summer for Cfa and warm summer for Cfb (PEEL et. al., 2007).

The region weather is influenced by a maritime polar (mP) air mass, a maritime tropical (mT) air mass and a continental tropical (cT) air mass. The maritime air masses are responsible for the moisture transport during the winter season, differing the South region of Brazil, with moister winters from the South East region that usually have a dry season during the winter. The precipitation indexes at the basin are steady over the year with total annual precipitations usually standing between 1200 mm and 2000 mm. The region climate is extremely influenced by the El Niño Southern Oscillation (ENSO). The warm phase of ENSO increases the Pacific Ocean water temperature in areas close to the Pacific coast of South America. This phenomenon may drastically increase the temperature and precipitation volume at the southern region of Brazil.

### 3.3 SELECTED STATIONS

The Brazilian Operator of the National Electricity System (Operador Nacional do Sistema Elétrico, ONS) currently provides series at the daily, monthly and annual scales for gauging stations at 153 hydroelectric power plants. The operation of plants and reservoirs cause significant changes for the river regimes. Thus, specific techniques, may be used in order to evaluate the natural flow of the river. By means of those techniques, the ONS obtains the naturalized streamflow series for all the aforementioned stations. The series are updated at the end of every year, by adding the naturalized streamflow values for the year

before. The considered time series for this study were all provided by the ONS at the monthly scale. By the time this research was taken, the provided monthly series were 87 years in length (1044 months), starting at January/1931 and ending by December/2017.

The selected gauging stations are six in number, downstream the six hydroelectric plants within the Iguaçu River, namely Foz do Areia, Segredo, Salto Santiago, Salto Osório, Salto Caxias and Baixo Iguaçu, adding up to a total of 7 024 MW of installed power and located in the Iguaçu River, as shown in Figure 10. The plants together compound about 7% of the Brazilian hydroelectric installed power.

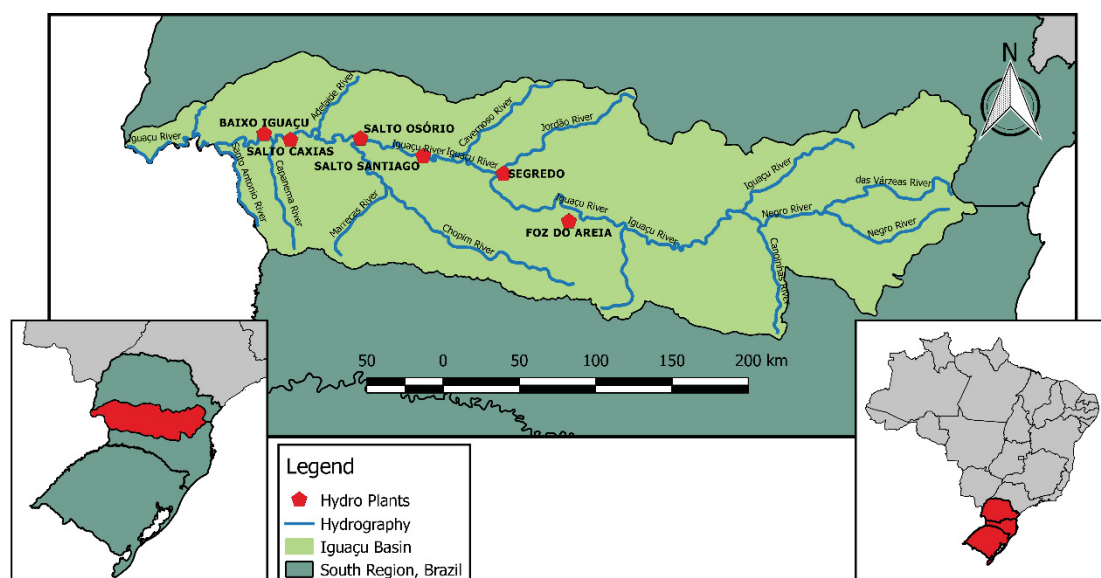


Figure 10. Map of the Iguaçu Basin

The Governador Bento Munhoz Rocha Neto power plant (i.e. Foz do Areia) started its operation in 1980 and has an installed power of 1 676 MW. The plant is located in the Iguaçu River, 5 km downstream the Areia River falls, at the municipality of Pinhão, Paraná. The delimited catchment has an average long-term flow of about 665 m<sup>3</sup>/s.

The Governador Ney Aminthas de Barros Braga power plant (i.e. Segredo) has 1 260MW of installed power and operates since 1992. It is located at the municipality of Mangueirinha, Paraná, in the Iguaçu River, 2 km upstream the Jordão River falls. The average long-term flow at the hydro plant is about 771 m<sup>3</sup>/s.

The following plant downstream Segredo in the Iguaçu River is the Salto Santiago power plant, which delimitates a sub-basin with an average long-term



flow of about 1 023 m<sup>3</sup>/s. The power plant is located in the municipality of Saudade do Iguaçu, Parana, and has an installed capacity of 1 420 MW.

Salto Osório is a hydroelectric plant within the Iguaçu River located in the municipality of Quedas do Iguaçu, Parana. The plant has 1 078 MW of installed power and the catchment it delimitates has an average long-term flow of about 1 071 m<sup>3</sup>/s.

Downstream Salto Osório and with 1240 MW of installed power is located the Governador José Richa power plant (i.e. Salto Caxias). The hydroelectric plant is located in the municipality of Capitão Leônidas Marques, Paraná and has an average long-term flow of about 1 375 m<sup>3</sup>/s.

Finally, with a foreseen installed power of 350 MW the Baixo Iguaçu power plant is currently under construction in the Iguaçu River at the municipality of Capitão Leônidas Marques, Paraná and at the plant the watercourse has an average long-term flow of about 1 486 m<sup>3</sup>/s. Supplementary details about the power plants used in this study are summarized in Table 2.

Table 2. Hydroelectric plants data

Municipality	Latitude	Longitude	Drainage Area (km <sup>2</sup> )	Average flow (m <sup>3</sup> /s)	Specific flow rate (m <sup>3</sup> /s/km <sup>2</sup> )	CV (%)
Pinhão, PR	26°00'34" S	51°40'00" W	30 127	665	0.0221	71.15%
Mangueirinha, PR	25°47'35" S	52°06'47" W	34 346	771	0.0224	70.49%
Saudade do Iguaçu, PR	25°37'04" S	52°36'48" W	43 852	1 023	0.0233	71.42%
Quedas do Iguaçu, PR	25°32'06" S	53°00'33" W	45 769	1 071	0.0234	71.40%
Capitão Leônidas Marques, PR	25°32'36" S	53°29'48" W	56 977	1 375	0.0241	71.16%
Capitão Leônidas Marques, PR	25°30'00" S	53°40'00" W	61 577*	1 486	0.0241*	71.17%

\* Estimated data considering the same Specific flow rate from the previous gauging station.

The main data were obtained from the files available at the ONS website. The Average flow refers to the average long-term flow at the monthly scale, the Specific flow rate is obtained by dividing the average flow by the drainage area and the CV is the mean of the coefficients of variation of the historical series for each month obtained by dividing the standard deviation of the month by its average flow. The mean CV analysis presents all above 70%. The series present similar behavior, July is the month that presents the highest CVs for all of them, from 91.16% at Gov. José Richa to 94.51% at Foz do Areia and March presents



the lowest CVs, from 52.16% at Foz do Areia to 55.43% at Baixo Iguaçu. These elevated values indicate a relatively sparse hydrologic regime.

Furthermore, the descriptive statistics of standard deviation ( $S_d$ ), skewness ( $g$ ), minimum, maximum and lag one correlation ( $\hat{\rho}_1$ ) for all the series are listed at Table 3. The elevated skewness suggests that a numerical transformation may be necessary for the series to tend to a normal behavior.

Table 3. Descriptive statistics

Series	$S_d$	$g$	$MIN (m^3/s)$	$MAX (m^3/s)$	$\hat{\rho}_1$
<b>Foz do Areia</b>	497	2.25	80	5150	0.492
<b>Segredo</b>	571	2.21	94	5893	0.497
<b>Salto Santiago</b>	772	2.36	116	8252	0.495
<b>Salto Osório</b>	808	2.33	119	8473	0.495
<b>Gov. José Richa</b>	1038	2.31	148	10798	0.488
<b>Baixo Iguaçu</b>	1122	2.31	160	11670	0.488

## 4 RESULTS AND DISCUSSION

### 4.1 PRELIMINARY ANALYSIS: STATIONARITY COMPARISON

As said earlier, the ARIMA model is not capable of dealing with trends in the series. Aiming to verify this hypothesis, the residuals of an ARMA model were submitted to the Mann-Kendal and Pettitt hypothesis-tests for stationarity, together with Sen slope estimation (KAHYA and KALAYCI, 2004). These analyses were performed for the six historical series at the yearly scale and their respective AR(1) residual series. As shown in Detzel (2015), the AR(1) model is adequate to model the Iguaçu river streamflow at the annual time scale.

The results of the trend tests are shown in Table 4, in which **year\_PT** and **p-value\_PT** refer respectively to the year of the break and p-value obtained by the Pettitt test, whereas **MK** and **p-value\_MK** correspond to the statistics and p-value of the Mann-Kendall test and the last column denotes the Sen Slope coefficient.

Table 4. Results of the trend tests

Series		year_PT	p-value_PT	MK	p-value_MK	Sen slope
Foz do Areia	History	1968	0,022	2,630	0,009	2,753
	Residual	1968	0,015	2,686	0,007	2,649
Segredo	History	1968	0,011	2,870	0,004	3,680
	Residual	1968	0,007	2,919	0,004	3,684
Salto Santiago	History	1968	0,019	2,801	0,005	4,833
	Residual	1968	0,012	2,896	0,004	4,888
Salto Osório	History	1968	0,017	2,824	0,005	5,050
	Residual	1968	0,011	2,949	0,003	5,147
Gov. José Richa	History	1968	0,014	2,855	0,004	6,787
	Residual	1968	0,007	3,055	0,002	6,928
Baixo Iguaçu	History	1968	0,014	2,955	0,003	7,604
	Residual	1968	0,009	3,064	0,002	7,812

The results acknowledge relevant trend and break for all historical series at a significance of 5%. Considering the results of the Man-Kendall test this is noticed at a significance of 1%. Moreover, the non-stationarity is also registered in the residuals, corroborating the hypothesis. Finally, for all tested series the year of the break obtained by the Pettitt test was maintained the same for both the historical and residual series, whereas the remaining parameters presented slight changes. The power plants are in the same cascade and are disposed on the

table in the same order they are in the watercourse, therefore, the flow gradually increases from one plant to the following.

At the Foz do Areia power plant the Pettitt test resulted in a p-value of 2.2% for the historical and 1.5% for the residual, indicating an increase in the significance of the non-stationarity. The same is registered for the Mann-Kendal test, in which the p-value decreases from 0.9% to 0.7%. The positive statistics of the Man-Kendal test indicates that both historical and residual series at Foz do Areia have increasing linear trends and the comparison between the Sen Slope coefficients indicates nearly the same linear trend. For illustration Figure 11 presents the two series and their respective linear trend.

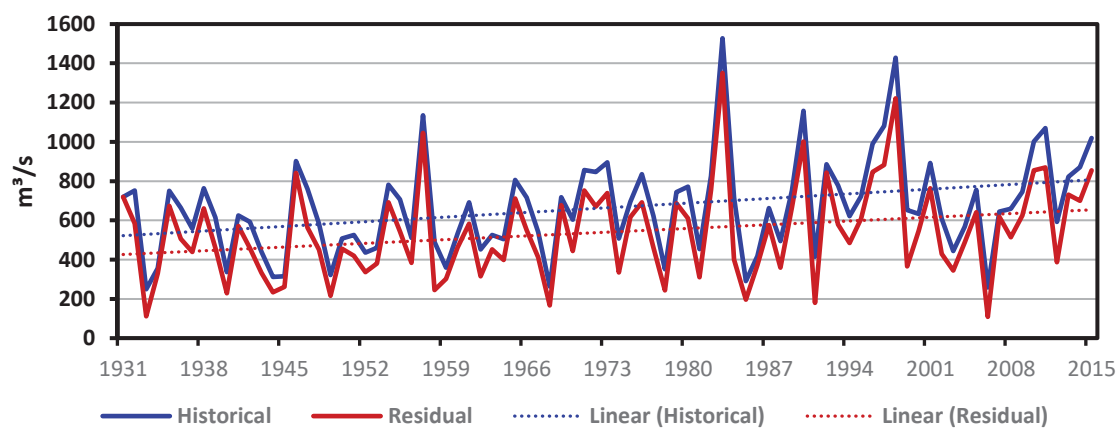


Figure 11. Historical and AR(1) residual series at Foz do Areia

Similar results were observed for the remaining series, as the behavior of the residuals in relation to the historical was equivalent to that observed at Foz do Areia. Additionally, they presented a slight and gradual increase in the Sen Slope coefficient while the Mann-Kendal and Pettitt p-values decreased in relation to Foz do Areia, meaning that the increasing trend is less intense and slightly less significant at Foz do Areia. In conclusion, the ARMA model is admittedly incapable of computing the trend of a series and the results reinforce that by demonstrating how this characteristic is kept by the residual series.

## 4.2 ARIMA MODEL

The preliminary analysis for the ARIMA model regards the iterative approach to model building of Box & Jenkins presented in Figure 3 (Page 14). The method suggests for the identification both graphic and theoretical approaches. Thus, for the graphic method the ACF and PACF analyzes are

required and for the theoretical identification the BIC was utilized. The ACF and PACF for the Foz do Areia series is presented in Figure 12.

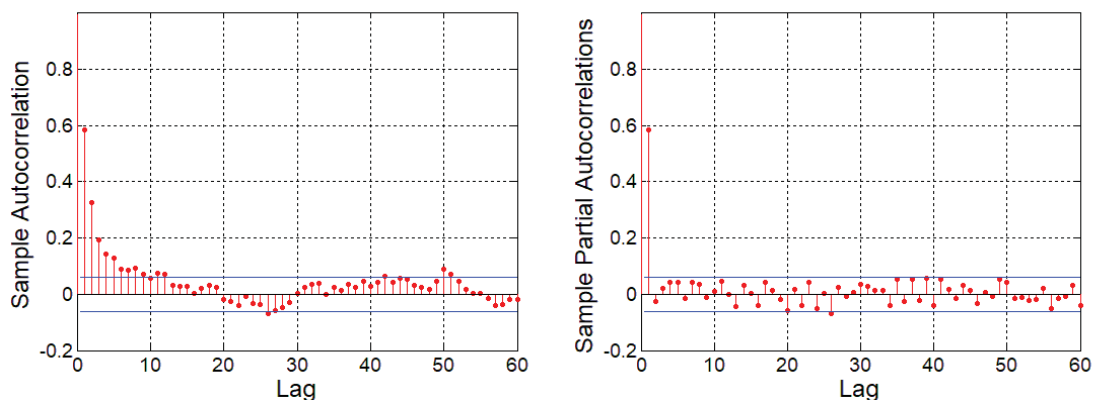


Figure 12. ACF and PACF at Foz do Areia

The graphics in Figure 12 show blue lines representing the significance limits, values placed in between those lines are considered statistically null. The exponential behavior presented by the ACF in addition to the sudden decay in the PACF right after the first lag suggest a pure autoregressive model of order one (ARIMA(1,0,0)). The same behavior is observed for the other five stations and the ACF and PACF plots for them are presented in Appendix A.1.

Table 5 presents the evaluated BIC values for the five ARIMA models considered in this research.

Table 5. BIC results for the ARIMA models tested

Series	Order of ARIMA ( $p,d,q$ ) model				
	(1,0,0)	(2,0,0)	(1,0,1)	(2,0,1)	(2,0,2)
Foz do Areia	<b>2533.30</b>	2538.14	2539.51	2542.96	2550.89
Segredo	<b>2510.98</b>	2515.85	2517.35	2520.28	2527.90
Salto Santiago	<b>2489.24</b>	2493.55	2495.12	2498.68	2506.08
Salto Osório	<b>2489.65</b>	2493.83	2495.39	2499.13	2506.43
Gov. José Richa	<b>2492.07</b>	2496.26	2497.83	2502.28	2509.17
Baixo Iguaçu	<b>2492.08</b>	2496.27	2497.84	2502.29	2509.18

The BIC results identify the ARIMA (1,0,0) as the proper model for all the six stations, corroborating with the conclusions for the ACF and PACF analysis. The following steps for the model building are the estimation of the parameters and the theoretical validation of the model. For the estimation, the maximum likelihood estimates method was used; finally for the validation the independence, homoscedasticity and normality the Portmanteau (LI and MCLEOD, 1981), Levene (BROWN and FORSYTHE, 1974) and Jarque-Bera (Ferreira, 2008) tests

were performed. The results for the estimation and validation for all the stations are listed in Table 6.

Table 6. ARIMA (1,0,0) parameters and theoretical validation p-values

Series	Parameter $\phi_1$	Portmanteau	p-values	
			Levene	Jarque-Bera
Foz do Areia	0.5843	0.302	0.867	0.004
Segredo	0.5961	0.131	0.978	0.003
Salto Santiago	0.6072	0.245	0.978	0.001
Salto Osório	0.6069	0.242	0.979	0.001
Gov. José Richa	0.6057	0.281	0.980	0.001
Baixo Iguaçu	0.6057	0.280	0.980	0.001

For a significance of 5%, the Portmanteau and Levene tests fail to reject the null hypothesis for all stations, meaning that in any case the residuals for an ARIMA (1,0,0) are considered to be independent and homoscedastic. The Jarque-Bera test, on the other hand, rejected the null hypothesis in all cases, meaning the residual series are not normally distributed. Further investigations on this matter were performed, by means of histogram plots (APPENDIX A.2). Results indicated that the histograms shapes closely resembled the typical bell-shaped normal curve. Hence, the option was to keep the ARIMA (1,0,0) model for the selected series.

In conclusion, the different series presented noticeably close results for the ARIMA model, a coherent behavior considering the stations are within the same river.

#### 4.3 ARIMA-ANN MODEL

The ANN was firstly trained for several different architectures for each station in order to determine the optimal among those tested. For the training, the residual series given by the selected AR(1) models were used, with 70% of the series being used for the estimation and 30% for the validation. Each combination between 1 to 18 neurons in the input layer and 1 to 36 neurons in the hidden layer were trained, resulting in 648 possible architectures per station. Afterwards, their respective RMSE were tested for both estimation and validation. Among these possibilities, the one that presented the smallest RMSE value for the validation was selected. Table 7 presents the optimal architectures to be used at each station and their respective RMSEs for estimation and validation.

Table 7. Optimal architectures and RMSE

Series	Architecture		RMSE		Regression $R^2$
	Input	Hidden Layer	Estimation	Validation	
Foz do Areia	18	32	0.0190	0.0170	0.9832
Segredo	18	32	0.0181	0.0231	0.9776
Salto Santiago	18	31	0.0165	0.0184	0.9762
Salto Osório	17	32	0.0222	0.0196	0.9794
Gov. José Richa	18	32	0.0233	0.0169	0.9867
Baixo Iguaçu	18	32	0.0173	0.0134	0.9839

The optimal architectures were the same for four of the six stations, with eighteen neurons in the input layer and thirty-six neurons in the hidden layer. Similar architectures were present for the other two stations, differing by one in the number of neurons in the hidden layer at Salto Santiago and in the input Layer at Salto Osório. The selected architectures were mainly those with the largest number of neurons, indicating that larger architectures would produce better results. The training, however, already present a good fit at these limits, as illustrated in Figure 13, in which the expected and the estimated outputs at Foz do Areia are plotted and the goodness of fit for the training is noticeable.

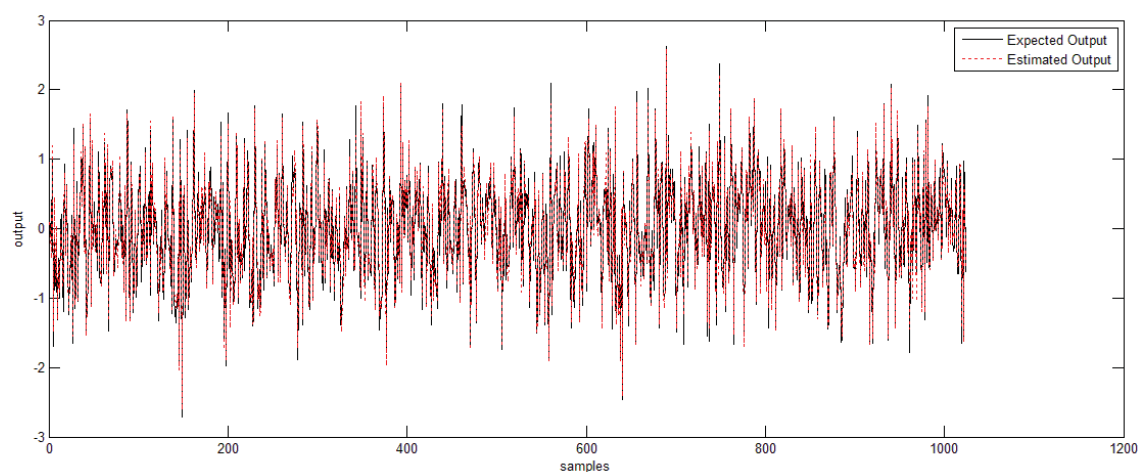


Figure 13. Expected and estimated outputs at Foz do Areia

The regression for the training also showed good results, with  $R^2$  values fluctuating around 0.98 at all of the stations. Figure 14 presents the plots for the regressions.

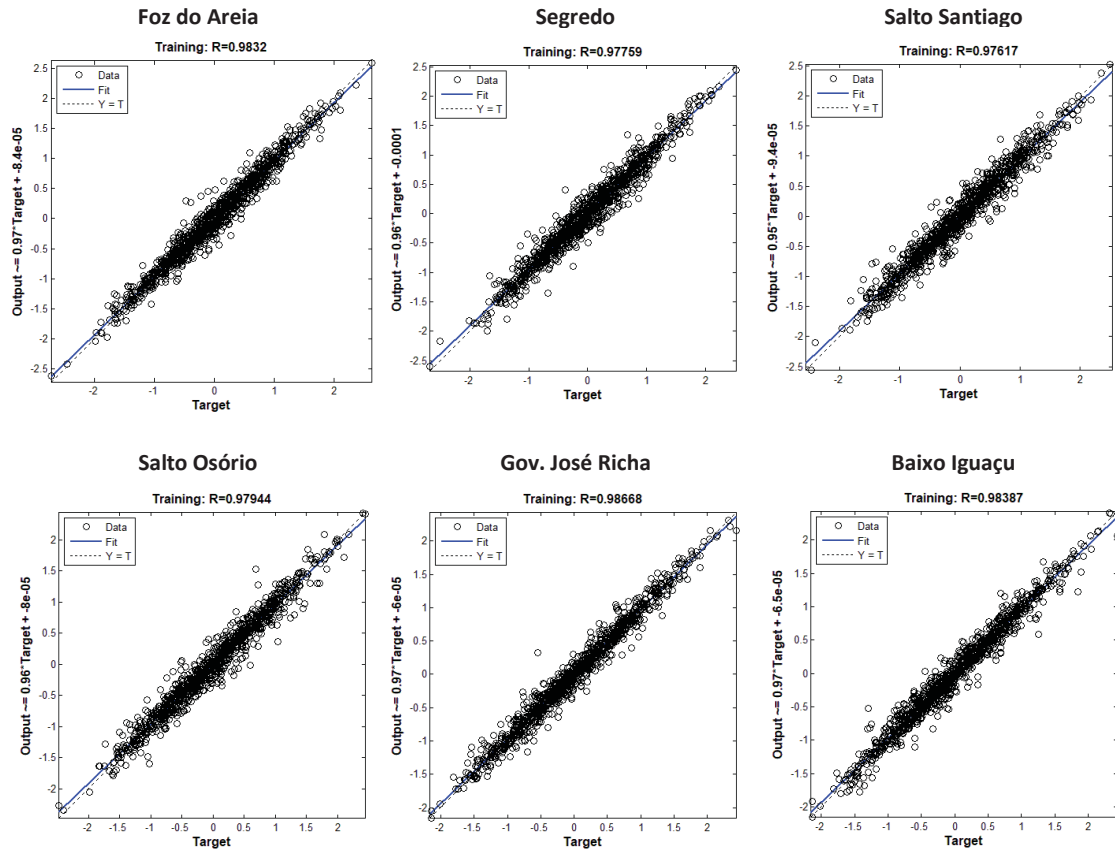


Figure 14. Training Regression

When using the trained networks for the synthetic generation, the first observed issue was regarding the processing time. Essentially, larger architectures lead the process to take longer to be processed, especially in comparison with the ARIMA model, the reasons however were not addressed by this study. For 1044 months (87 years) long series, the computer was spending approximately 25 seconds per generated series. Considering a total of 1000 synthetic series for each of the 6 stations to be generated, the processing time would total almost 42 hours for producing the 6000 series. Before running the computer for that long, only 10 synthetic series for each station were generated in order to verify the behavior of the synthetic series. Figure 15 presents the residual series from the ARIMA model  $a_t$  and the synthetic series of residuals  $a_t^*$  produced by a MLP with 18 neurons in the input layer and 32 neurons on the hidden layer. The figure shows the results for Foz do Areia, nevertheless, the behavior was similar in all the stations.

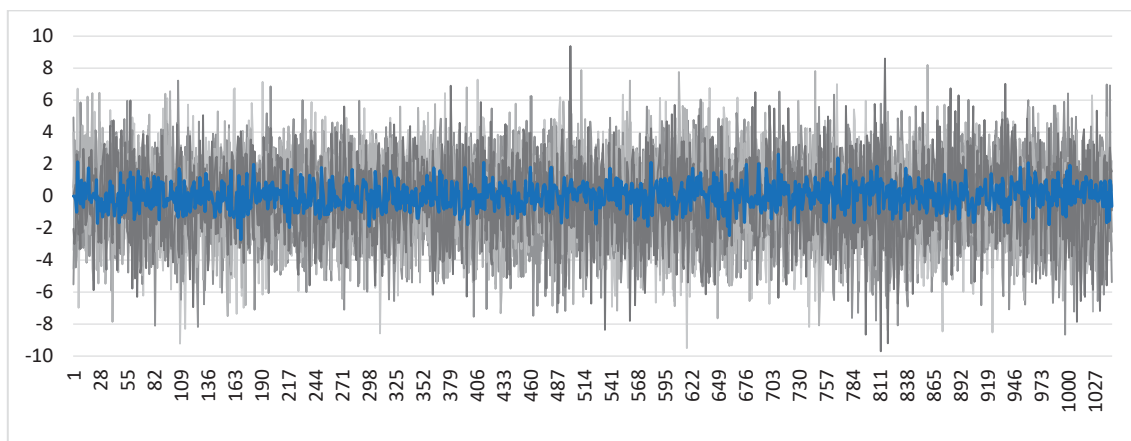


Figure 15. Synthetic series of residuals – Foz do Areia

In a quick graphic analysis, it is noticeable that the synthetic series of residuals present a similar behavior to the one observed in the original series, with mean values fluctuating around the zero. The synthetic series, however, present much lower minimum values and higher maximum values, incurring in a higher standard deviation. When returning the series to the ARIMA formulation the synthetic streamflow series are obtained, the plots of the historical series  $z_t$  and the synthetic series  $z_t^*$  at Foz do Areia are shown in Figure 16. The other stations have produced similar results.

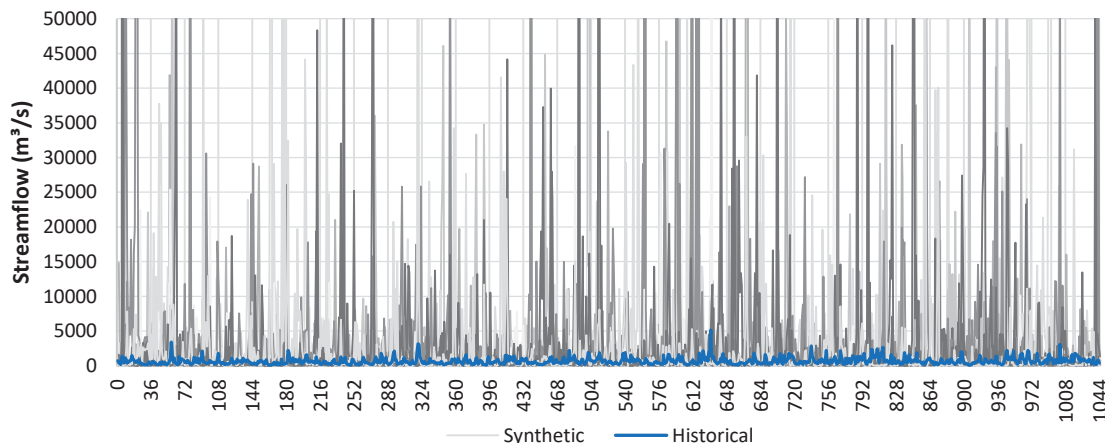


Figure 16. Synthetic streamflow series – Foz do Areia 1

The graphic shows far higher maximum values for the synthetic streamflow than those registered. In fact, the maximum synthetic values register an average of more than 100 times bigger than the maximum registered. Moreover, the maximum single value produced was almost 500 times bigger than the maximum registered. Therefore, no theoretical test was needed to conclude the model results were not good. In spite of those unlikely values, the behavior similarities observed in the Figure 15 graphical analysis, lead to the assumption



that it was a matter of calibration. Thus, a trial and error analysis was performed in search for the most suitable architecture. Since it was impracticable to produce 6000 series with each architecture, for each check would take almost two days, the tests were made by generating 10 series for each architecture for each station.

Some initial trials revealed that more complex networks, with more neurons and far better fit would produce worse results and even bigger maximum values. Too simple a network, on the other hand, would lack meaningful information and be incapable of properly reproduce the series behavior. The process was repeated for all the six stations and some fine tuning led to suitable architectures for synthetic streamflow generation at these stations. The weights were different for each station, but the main structure was the same and the all six series were modeled by a 12 neuron input layer 18 neuron hidden layer MLP. The result for the 10 synthetic series generation at Foz do Areia is presented in Figure 17.

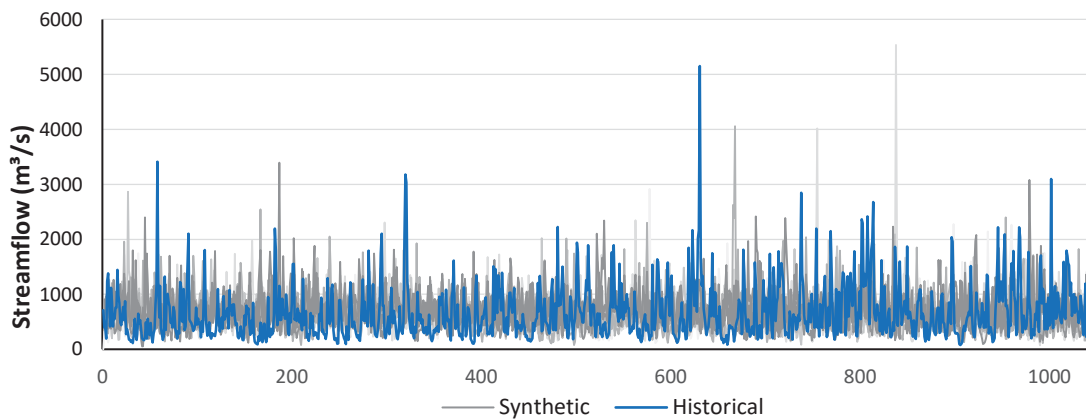


Figure 17. Synthetic streamflow series – Foz do Areia 2

The 12-18-1 architecture is further used to generate 1000 series to be compared with the single ARIMA model.

#### 4.4 COMPARATIVE ANALYSIS

Hybrid and single models were used to generate 1000 streamflow series with 1044 elements each, at the monthly scale. This section presents a comparison between the synthetic series generated by single ARIMA and ARIMA-ANN models. The short-term statistics and their uncertainties are presented in Table 8, noting that the results for the synthetic series refer to the

mean of the statistics for all the generated scenarios and the values between parentheses correspond to the uncertainties. Results for the synthetic series statistics, when compared with those from the historical series present a fair adherence for both models.

The single ARIMA model performed better in terms of mean ( $\hat{\mu}$ ), standard deviation ( $\hat{\sigma}$ ) and maximum values, whilst the hybrid model better represented the skewness ( $\hat{\xi}$ ) and the minimum values. Specifically for the skewness ( $\hat{\xi}$ ), the ARIMA-ANN model presented a considerably better performance over the single ARIMA. Moreover, one may notice that for statistics in which the single ARIMA model performs better, the hybrid model presents a lower uncertainty and where the ARIMA-ANN model overcomes, it presents a higher uncertainty.

Table 8. Short-term statistics and uncertainty

Series	Method	$\hat{\mu}$ ( $m^3/s$ )	$\hat{\sigma}$ ( $m^3/s$ )	$\hat{\xi}$	MIN ( $m^3/s$ )	MAX ( $m^3/s$ )
<b>Foz do Areia</b>	Historical	665	497	2.25	80	5150
	ARIMA	668 (29.4)	529 (55.0)	3.07 (1.2)	50 (14.1)	5645 (1880)
	ARIMA-ANN	674 (20.3)	352 (43.7)	2.33 (1.5)	84 (25.0)	3754 (1592)
<b>Segredo</b>	Historical	771	571	2.21	94	5893
	ARIMA	773 (33.1)	609 (64.6)	3.03 (1.3)	60 (16.7)	6432 (2350)
	ARIMA-ANN	786 (24.3)	411 (51.1)	2.33 (1.5)	98 (29.4)	4379 (1847)
<b>Salto Santiago</b>	Historical	1023	772	2.36	116	8252
	ARIMA	1029 (46.8)	826 (85.2)	3.16 (1.0)	81 (21.8)	8883 (2749)
	ARIMA-ANN	1049 (33.7)	561 (70.9)	2.40 (1.5)	129 (39.0)	6011 (2538)
<b>Salto Osório</b>	Historical	1071	808	2.33	119	8473
	ARIMA	1076 (50.3)	861 (94.0)	3.06 (1.1)	83 (23.6)	9062 (3006)
	ARIMA-ANN	1098 (35.2)	588 (74.1)	2.40 (1.5)	135 (40.8)	6296 (2652)
<b>Gov. José Richa</b>	Historical	1375	1038	2.31	148	10798
	ARIMA	1383 (63.3)	1114 (114.0)	3.08 (1.2)	112 (31.3)	11766 (3930)
	ARIMA-ANN	1409 (45.1)	761 (94.2)	2.42 (1.5)	178 (52.6)	8136 (3376)
<b>Baixo Iguaçu</b>	Historical	1486	1122	2.31	160	11670
	ARIMA	1491 (66.8)	1202 (123.7)	3.12 (1.0)	121 (31.2)	12723 (3900)
	ARIMA-ANN	1523 (48.8)	823 (101.8)	2.42 (1.5)	193 (56.8)	8793 (3648)

$\hat{\mu}$  – mean;  $\hat{\sigma}$  – standard deviation;  $\hat{\xi}$  – skewness.

The statistics of monthly average and standard deviation at Foz do Areia are presented in Figure 18, the other series displayed similar behavior and therefore are presented in APPENDIX A.3. The ARIMA-ANN model noticeably underestimates the standard deviations, whereas the single ARIMA behaves rather similar to that from the historical series. The monthly average does not present much of a difference, the hybrid model performed a little better for June, August and September, although the ARIMA model better preserved the statistics overall.

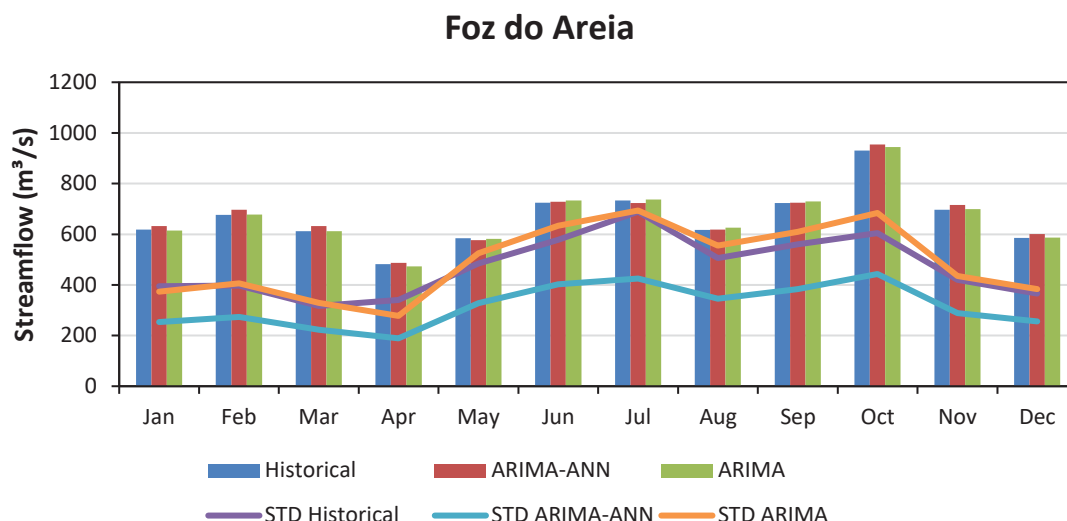


Figure 18. Monthly statistics – Foz do Areia. Bars indicate means, and lines indicate standard deviations.

The comparison between the autocorrelation functions at Foz do Areia (Figure 19) indicate a faster decay for the ARIMA-ANN model, while the historical series presents statistical null correlations from lag 6 on, the ARIMA-ANN reaches the limit at lag 3 and the single ARIMA at lag 5. At further lags the hybrid better approaches the historical, especially at lags 11 and 12, in which the correlations become significant again and the hybrid coincides with historical. Moreover, the proposed model presents positive autocorrelation at lags 24, 36, 48, and 60, and negative at lags 6, 18, 30, 42, and 54, demonstrating some sort of seasonality that not represents the historical series. The ACF at the other stations present similar behavior and are shown in APPENDIX A.4.

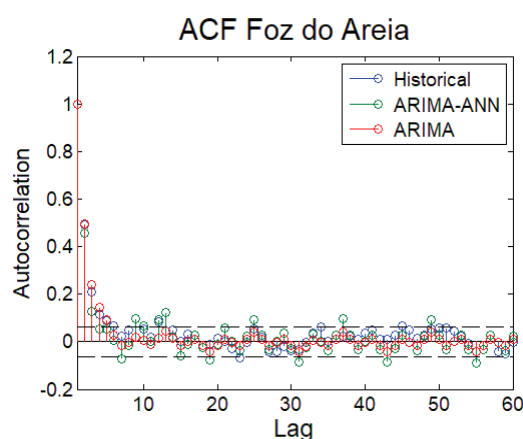


Figure 19. Autocorrelation Functions – Foz do Areia

The long-term statistics are presented in Table 9, with the uncertainties regarding these statistics shown between parentheses. The results indicate that the single ARIMA mainly prevails over the ARIMA-ANN regarding the long-term statistics, with the exception for the average run period, in which the proposed

model presents better results. The uncertainty however, was lower for the hybrid model in comparison to the classic for the vast majority of long-term statistics, with the number of runs at Segredo as the only exception.

Table 9. Long-term statistics and uncertainty

Series	Method	$n_A$	$\bar{\Lambda}$	$\Lambda_{max}$	$\bar{\Lambda}$ ( $m^3/s$ )	$\Lambda_{max}$ ( $m^3/s$ )	$\bar{\Delta}$ ( $m^3/s$ )	$\Delta_{max}$ ( $m^3/s$ )
Foz do Areia	Historical	113	5	31	2033	9202	1378	7224
	ARIMA	110 (6.5)	6 (0.39)	24 (5.4)	2176 (185)	8589 (1954)	1395 (426)	7775 (2388)
	ARIMA-ANN	122 (6.2)	5 (0.25)	18 (4.4)	2044 (119)	7375 (1617)	506 (152)	4413 (1547)
Segredo	Historical	111	6	31	2441	10193	1965	9731
	ARIMA	109 (6.1)	6 (0.37)	25 (5.3)	2572 (203)	10136 (2171)	1654 (439)	9138 (2668)
	ARIMA-ANN	121 (6.2)	5 (0.26)	19 (4.5)	2421 (144)	8800 (1966)	618 (188)	5330 (1868)
Salto Santiago	Historical	110	6	31	3242	13238	2899	14581
	ARIMA	107 (6.2)	6 (0.40)	25 (5.3)	3475 (291)	13646 (2861)	2302 (670)	12536 (3853)
	ARIMA-ANN	120 (6.0)	5 (0.26)	19 (4.6)	3282 (201)	12013 (2703)	900 (276)	7557 (2636)
Salto Osório	Historical	111	6	31	3374	13817	3046	15521
	ARIMA	107 (6.1)	6 (0.38)	26 (5.8)	3636 (292)	14293 (3315)	2440 (749)	13286 (4110)
	ARIMA-ANN	120 (6.0)	5 (0.26)	19 (4.5)	3439 (210)	12543 (2756)	942 (289)	7911 (2758)
Gov. José Richa	Historical	116	5	33	4320	20374	3946	21071
	ARIMA	107 (6.2)	6 (0.40)	26 (5.7)	4678 (385)	18367 (4178)	3181 (869)	17395 (4912)
	ARIMA-ANN	121 (5.9)	5 (0.26)	19 (4.5)	4439 (278)	16158 (3546)	1213 (370)	10161 (3534)
Baixo Iguaçu	Historical	116	5	33	4669	22022	4264	22765
	ARIMA	107 (6.3)	6 (0.39)	26 (6.3)	5037 (409)	19847 (4714)	3350 (949)	18459 (5507)
	ARIMA-ANN	121 (5.9)	5 (0.26)	19 (4.5)	4799 (301)	17467 (3838)	1311 (400)	10981 (3820)

$n_A$  – number of runs;  $\bar{\Lambda}$  – average run period;  $\Lambda_{max}$  – maximum run period;  $\bar{\Delta}$  – average deficit;  $\Delta_{max}$  – maximal accumulated deficit.

Generally, the proposed model reduces the uncertainty and shows a significant improvement in terms of skewness. On the other hand, it underestimates the standard deviation and the maximums. The lower values for standard deviation, in addition to the seasonal pattern observed in the ACF indicate that the neural network detects a well-established seasonality that does not reflect the Iguaçu River regime. Considering the long-term statistics, the model is fair on average, but underestimates the maximal values.

## 5 CONCLUSION

An alternative approach on the use of Neural Networks in hydrology was considered in this research. The purpose of this work was to verify and evaluate the efficiency improvement for a hybrid ARIMA-ANN in comparison with an ARIMA model for synthetic streamflow generation, relying on the premise that the neural network would address the non-linear portion of the river regime, whilst the ARIMA part would compute the persistence of the series.

The model building incurred in some challenging tasks. Firstly, regarding the coupling scheme, since there are many researches with hybrid models similar to this but no consensus on how to join the two models in one. The ANN should filter what the ARIMA model was not able to, thus the option for modelling the ARIMA residual with the ANN, for the residual contains all the information not computed by the ARIMA. In the sequence it was noticed that optimizing the architecture by means of the RMSE and the coefficient of determination ( $R^2$ ) wouldn't improve the results for the synthetic streamflow generation. Therefore, followed an extensive search for the most suitable architecture. One should attempt to the fact that this search must be repeated for each station.

The main problem with the model was the processing time. One synthetic series alone consumes a few seconds to be generated. Taking into account the number of synthetic series to be generated and the number of different stations, the model would take a considerable amount of time to be processed, making the model impracticable for most operational studies. A possible reason to that might be the usage of the Matlab neural network toolbox, although the computational cost was not considered in this research and this issue could be further addressed.

The proposed model was especially good in accessing the minimal streamflow values, despite underestimating the maximums. Thus, it could be suitable for studies of drought events, in which the lower values prevail over the higher. Moreover, there was some improvement in reducing the uncertainty, which can be due to considering the non-linearity inherent to the series.

The objectives proposed in the introduction of this research were met and the model is appropriate for the synthetic streamflow generation within the Iguaçu

Basin, although the implementation and processing times should be previously considered. That said, future research can derive from this:

- The underestimation of the standard deviation is possibly due to the underestimation of the maximums. This problem however, can be a matter of calibration and should be further investigated and improved.
- The synthetic series generation demands an elevated quantity of series to be generated. Therefore, a process that takes a few seconds can take several hours to be completed when repeated many times. The forecasting however, require only one short series, time dependent on the historical series, meaning that the processing time would not be as much of a problem.
- The issue regarding the elevated implementation and processing time could be solved by optimization techniques.
- One way to extend the model to the multivariate analysis is by using data from more stations as inputs and by producing more outputs. This should directly reflect in the computational cost but could be further investigated.

In conclusion, this study can further be extended and improved in diverse ways, bringing new possibilities of studies and expanding the knowledge in synthetic hydrology.

## REFERENCES

- ABRAHART, R. J.; SEE, L. Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments. **Hydrological Processes**, v. 14, n. 11–12, p. 2157–2172, 2000. Available at: <<http://doi.wiley.com/10.1002/1099-1085%2820000815/30%2914%3A11/12%3C2157%3A%3AAID-HYP57%3E3.0.CO%3B2-S>>.
- ADAM, K. N.; COLLISCHONN, W.. Análise dos impactos de mudanças climáticas nos regimes de precipitação e vazão na bacia hidrográfica do rio Ibicuí. **Revista Brasileira de Recursos Hídricos**, v. 18, n. 3, p. 69-79, 2013.
- ADELOYE, A. J. Multiple Linear Regression and Artificial Neural Networks Models for Generalized Reservoir Storage–Yield–Reliability Function for Reservoir Planning. **Journal of Hydrologic Engineering**, v. 14, n. 7, p. 731–738, 2009.
- AHMED, J. A.; SARMA, A. K. Artificial neural network model for synthetic streamflow generation. **Water Resources Management**, v. 21, n. 6, p. 1015–1029, 2007.
- AKSOY, H.; DAHAMSHEH, A. Markov chain-incorporated and synthetic data-supported conditional artificial neural network models for forecasting monthly precipitation in arid regions. **Journal of Hydrology**, v. 562, n. May, p. 758–779, 2018. Elsevier. Available at: <<https://doi.org/10.1016/j.jhydrol.2018.05.030>>.
- ALLAWI, M. F.; EL-SHAFIE, A. Utilizing RBF-NN and ANFIS Methods for Multi-Lead ahead Prediction Model of Evaporation from Reservoir. **Water Resources Management**, v. 30, n. 13, p. 4773–4788, 2016. Water Resources Management. Available at: <<http://dx.doi.org/10.1007/s11269-016-1452-1>>.
- BATISTA, A. L., FREITAS JR., S. A. de, DETZEL, D. H. M., MINE, M. R. M. FILL, H. D. O. A., FERNANDES, C. KAVISKI, E. Verificação da estacionariedade de séries hidrológicas no Sul-Sudeste do Brasil. In SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS, 18., 2009. Campo Grande. **Anais...** Porto Alegre: ABRH, 2009, p. 1-19.
- BAYER, D. M.; CASTRO, N. M. R. (2012). Modelagem e previsão de vazões médias mensais do rio Potiribu utilizando modelos de séries temporais. **Revista Brasileira de Recursos Hídricos**, 17(2), p. 229-239, 2012.
- BORGOMEIO, E.; FARMER, C. L.; HALL, J. W. Numerical rivers: A synthetic streamflow generator for water resources vulnerability assessments. **Water Resources Research**, v. 51, n. 7, p.5382- 5405, 2015.
- BORMANN, H., PINTER, N., ELFERT, S. Hydrological signatures of flood trends on German rivers: Flood frequencies, flood heights and specific stages. **Journal of Hydrology**, v. 404, n. 1-2, p. 50–66, 2011.

BOX, G. E. P., JENKINS, G. M., REINSEL, G. C. **Time Series Analysis Forecasting and Control** 4<sup>a</sup> ed. New Jersey: John Wiley & Sons, 2008.

CARNEIRO, T. C.; FARIAS, C. A. S. Otimização estocástica implícita e redes neurais artificiais para auxílio na operação mensal dos reservatórios Coremas - Mãe d' Água. **Revista Brasileira de Recursos Hídricos**, v. 18, n. 4, p. 115–124, 2013.

COULIBALY, P.; ANCTIL, F.; ASCE, M.; BOBÉE, B. Multivariate reservoir inflow forecasting using temporal neural networks. **Journal of Hydrologic Engineering**, v. 6, p. 367–376, 2001.

CHOW, V. T. Handbook of applied hydrology, **MacGraw-Hill Book Co.**, 1964.

DETZEL, D. H. M. **Modelagem de séries hidrológicas : uma abordagem de múltiplas escalas temporais**. 218 p. Thesis (Doctorate at Water Resources and Environmental Engineering) – Setor de Ciências Exatas e Setor de Tecnologia, Universidade Federal do Paraná, Curitiba, 2015.

DETZEL, D. H. M.; MINE, M. R. M. Comparison between Deseasonalized Models for Monthly Streamflow Generation in a Hurst–Kolmogorov Process Framework. **Journal of Hydrologic Engineering**, v. 22, n. 4, p. 05016040, 2016. Available at: <<http://ascelibrary.org/doi/10.1061/%28ASCE%29HE.1943-5584.0001488>>.

DETZEL, D. H. M.; MINE, M. R. M.; BESSA, M. R.; BLOOT, M. Cenários Sintéticos de Vazões para Grandes Sistemas Hídricos Através de Modelos Contemporâneos e Amostragem. **Revista Brasileira de Recursos Hídricos**, v. 19, n. 1, p. 17–28, 2014.

DETZEL, D. H. M., BESSA, M. R., VALLEJOS, C. A. V, SANTOS, A. B., THOMSEN, L. S. Estacionariedade das Afluências às Usinas Hidrelétricas Brasileiras. **Revista Brasileira de Recursos Hídricos**, v. 16, n. 3, p. 95–111, 2011.

DETZEL, D. H. M.; MEDEIROS, L.; OENING, A. P.; MARCILIO, D. C.; TOSHIOKA, F. Acerca da quantidade de simulações estocásticas de vazão no contexto do planejamento energético. **Revista Brasileira de Energia**, v. 22, n. 2, p. 21-32, 2016.

DOOGE, J. C. I., Linear theory of hydrologic systems, Technical Bulletin No. 1468, **U.S. Department of Agriculture**, 1973.

FARUK, D. Ö. A hybrid neural network and ARIMA model for water quality time series prediction. **Engineering Applications of Artificial Intelligence**, v. 23, n. 4, p. 586–594, 2010.

FERREIRA, D. F. **Estatística Multivariada**. Lavras: UFLA, 2008.

FERREIRA, L. F. N.; MINE, M. R. M.; FILL, H. D.; MACHADO, F. W. Monthly rainfall-modeling using artificial neural networks in the context of Claris LPB Project. , p. 14, 2011.



FILL, H. D. Análise da estacionariedade das vazões do rio iguaçu em União da Vitória. In: SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS, 19., 2011. Maceió. **Anais...** Porto Alegre: ABRH, 2011.

FLEMING, S. W., WEBER, F. A. Detection of long-term change in hydroelectric reservoir inflows: Bridging theory and practice. **Journal of Hydrology**, v. 470-471, p. 36–54, 2012.

GALVÃO, C. DE O.; VALENÇA, M. J. S.; VIEIRA, V. P. P. B.; et al. **Sistemas inteligentes: aplicações a recursos hídricos e sistemas ambientais**. 1st ed. Porto Alegre: ABRH, 1999.

GUIMARÃES, R. C.; SANTOS, E. G. Principles of stochastic generation of hydrologic time series for reservoir planning and design: Case study. **Journal of Hydrologic Engineering**, v. 16, n. 11, p. 891–898, 2011.

HAGHIABI, A. H. Modeling River Mixing Mechanism Using Data Driven Model. **Water Resources Management**, v. 31, n. 3, p. 811–824, 2017. Water Resources Management.

HALTINER, J. P., SALAS, J. D. Development and testing of a multivariate, seasonal ARMA(1,1) model. **Journal of Hydrology**, v. 104, p. 247–272, 1988.

HAYKIN, S. **Neural networks: a comprehensive foundation**. 9º ed. Pearson Education, 1999.

HIPEL, K. W.; MCLEOD, A. I. 1994-Time-chapter 14.pdf. **Time Series Modelling of Water Resources and Environmental Systems**, 1994.

HIRSCH, R. M. Synthetic hydrology and water supply reliability. **Water Resources Research**, v. 15, n. 6, p.1603-1615, 1979.

JACKSON, B. B. The use of streamflow models in planning. **Water Resources Research**, 11(1), p. 54 – 63, 1975.

KAHYA, E.; KALAYCI, S. Trend analysis of streamflow in Turkey. **Journal of Hydrology**, v. 289, n. 1–4, p. 128–144, 2004.

KASIVISWANATHAN, K. S.; HE, J.; SUDHEER, K. P.; TAY, J. H. Potential application of wavelet neural network ensemble to forecast streamflow for flood management. **Journal of Hydrology**, v. 536, p. 161–173, 2016. Elsevier B.V. Available at: <<http://dx.doi.org/10.1016/j.jhydrol.2016.02.044>>.

KAVISKI, Eloy. **Solução de problemas de fenômenos de transporte pelo método de Monte Carlo**. 330 p. Thesis (Doctorate at Numerical Methods for Engineering) – Setor de Ciências Exatas e Setor de Tecnologia, Universidade Federal do Paraná, Curitiba, 2006.

KELMAN, J. Modelos Estocásticos no Gerenciamento de Recursos Hídricos. **Modelos para Gerenciamento de Recursos Hídricos I**. São Paulo: Nobel/ABRH. 1987, cap. 4.

KHASHEI, M.; BIJARI, M. An artificial neural network (p, d, q) model for timeseries forecasting. **Expert Systems with Applications**, v. 37, n. 1, p. 479–489, 2010. Elsevier Ltd. Available at: <<http://dx.doi.org/10.1016/j.eswa.2009.05.044>>.

LIANG, L., LI, L., LIU, Q. Precipitation variability in Northeast China from 1961 to 2008. **Journal of Hydrology**, v. 404, n. 1-2, p. 67–76, 2011.

MACHADO, F.; MINE, M.; KAVISKI, E.; FILL, H. Monthly rainfall–runoff modelling using artificial neural networks. **Hydrological Sciences Journal**, v. 56, n. 3, p. 349–361, 2011. Available at: <<http://www.tandfonline.com/doi/abs/10.1080/02626667.2011.559949>>.

MAIER, H. R.; JAIN, A.; DANDY, G. C.; SUDHEER, K. P. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. **Environmental Modelling and Software**, v. 25, n. 8, p. 891–909, 2010.

MARTINI FILHO, L. R.; DETZEL, D. H. M.; PLOSZAI, R.; BESSA, M. R.; DE GEUS, K. Paramétricos Streamflow Synthetic Series Generation Through Parametric Models. **XXII Simpósio brasileiro de Recursos hídricos**, Florianópolis, Brazil, 2017.

MATALAS, N. C. Mathematical assessment of synthetic hydrology. **Water Resources Research**, v. 3, n. 4, p. 937–945, 1967.

NEIRA, K. L. Curvas de Regularização para Reservatórios Parcialmente Cheios e Confiabilidade Constante. 281 f. **Universidade Federal do Paraná**, Curitiba, 2005.

NGUYEN-KY, T.; MUSHTAQ, S.; LOCH, A.; et al. Predicting water allocation trade prices using a hybrid Artificial Neural Network-Bayesian modelling approach. **Journal of Hydrology**, 2017. Elsevier B.V. Available at: <<https://doi.org/10.1016/j.jhydrol.2017.11.049>>.

OCHOA-RIVERA, J. C. Prospecting droughts with stochastic artificial neural networks. **Journal of Hydrology**, v. 352, n. 1–2, p. 174–180, 2008.

ÖMER FARUK, D. A hybrid neural network and ARIMA model for water quality time series prediction. **Engineering Applications of Artificial Intelligence**, v. 23, n. 4, p. 586–594, 2010.

PATSKOSKI, J.; SANKARASUBRAMANIAN, A. Improved reservoir sizing utilizing observed and reconstructed streamflow within a Bayesian combination framework. **Water Resources Research**, v. 51, n. 7, p. 5677–5697, 2015.

PEEL, M. C., Finlayson, B. L., and McMahon, T. A.: Updated world map of the Köppen-Geiger climate classification, **Hydrol. Earth Syst. Sci.**, 11, 1633–1644, <https://doi.org/10.5194/hess-11-1633-2007>, 2007.

PÉRICO, G. Avaliação Estocástica da Influência de Reservatórios na Expansão de um Sistema Hidrelétrico. **Universidade Federal do Paraná**, Curitiba, 2014.

PRADA-SARMIENTO, F.; OBREGÓN-NEIRA, N. Forecasting of Monthly Streamflows Based on Artificial Neural Networks. **Journal of Hydrologic Engineering**, v. 14, n. 12, p. 1390–1395, 2009. Available at: <<http://ascelibrary.org/doi/10.1061/%28ASCE%291084-0699%282009%2914%3A12%281390%29>>.

RAUDKIVI, A. J. Hydrological Modelling and Water Resources Systems in Hydrology: An Advanced Introduction to Hydrological Processes and Modelling. **University of Auckland** – New Zealand, p. 347 – 379, 1979.

ROUGÉ, C., GE, Y., CAI, X. Detecting gradual and abrupt changes in hydrological records. **Advances in Water Resources**, v. 53, n. 33–44, 2013.

SALAS, J. D.; DELLEUR, J. W.; YEVJEVICH, Y.; LANE, W. L. Applied modeling of hydrologic time series. Chelsea, MI, U.S.A. **Water Resources Publications**, 484 p., 1980.

SAMMEN, S. S.; MOHAMED, T. A.; GHAZALI, A. H.; EL-SHAFIE, A. H.; SIDEK, L. M. Generalized Regression Neural Network for Prediction of Peak Outflow from Dam Breach. **Water Resources Management**, v. 31, n. 1, p. 549–562, 2017. *Water Resources Management*. Available at: <<http://dx.doi.org/10.1007/s11269-016-1547-8>>.

SCHMID, B. H.; ASCE, M.; KOSKIAHO, J. Artificial Neural Network Modeling of Dissolved Oxygen in a Wetland Pond: The Case of Hovi, Finland. , v. 11, n. April, p. 188–192, 2006.

SCHWARTZ, G. Estimating the Dimension of a Model. **The Annals of Mathematical Statistics**, v. 6, n. 2, p. 461–464, 1978.

SEYEDASHRAF, O.; MEHRABI, M.; AKHTARI, A. A. Novel approach for dam break flow modeling using computational intelligence. **Journal of Hydrology**, v. 559, p. 1028–1038, 2018. Elsevier B.V. Available at: <<https://doi.org/10.1016/j.jhydrol.2018.03.001>>.

SILVA, F. E.; NAGHETTINI, M.; FERNANDES, W. Avaliação bayesiana das incertezas nas estimativas dos parâmetros de um modelo chuva-vazão conceitual. **Revista Brasileira de Recursos Hídricos**, [s.l.], v. 19, n. 4, p. 148–159, 2014.

SHAO, Q.; WONG, H.; LI, M.; IP, W. Streamflow forecasting using functional-coefficient time series model with periodic variation. **Journal of Hydrology**, v. 368, n. 1–4, p. 88–95, 2009. Elsevier B.V. Available at: <<http://dx.doi.org/10.1016/j.jhydrol.2009.01.029>>.

SHOAIB, M.; SHAMSELDIN, A. Y.; MELVILLE, B. W.; KHAN, M. M. A comparison between wavelet based static and dynamic neural network approaches for runoff prediction. **Journal of Hydrology**, v. 535, p. 211–225, 2016. Elsevier B.V. Available at: <<http://dx.doi.org/10.1016/j.jhydrol.2016.01.076>>.

SOUZA, R. C., CAMARGO, M. E. **Análise e previsão de Séries Temporais: os modelos ARIMA**. 2 a ed., Rio de Janeiro: Regional, 2004.

SPECHT, D. F. A general regression neural network. **Neural Networks, IEEE Transactions on**, v. 2, n. 6, p. 568–576, 1991.

SRIVASTAV, R. K., SRINIVASAN, K., and SUDHEER, K. P. “Simulation-optimization framework for multi-season hybrid stochastic models.” **Journal of Hydrology**, v.404 n. 3-4, p. 209–225, 2011.

SUDHEER, K. P. Knowledge Extraction from Trained Neural Network River Flow Models. **Journal of Hydrologic Engineering**, v. 10, n. 4, p. 264–269, 2005.

SUDLER, C. E. Storage required for the regulation of stream flow, **Trans. Am. Soc. Civ. Eng.**, 91, p. 622–660, 1927.

TAGHI SATTARI, M.; YUREKLI, K.; PAL, M. Performance evaluation of artificial neural network approaches in forecasting reservoir inflow. **Applied Mathematical Modelling**, v. 36, n. 6, p. 2649–2657, 2012. Elsevier Inc. Available at: <<http://dx.doi.org/10.1016/j.apm.2011.09.048>>.

THOMAS, B. E. Climatic Fluctuations and Forecasting of Streamflow in the Lower Colorado River Basin. **Journal of the American Water Resources Association**, v. 43, n. 6, p. 1550–1569, 2007.

THOMAS, H. A., FIERING, M. B. Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. In Maass A. et al. (Org.), **Design of Water Resources Systems**. Cambridge: Harvard University Press. p. 459-493, 1962.

TODINI, E. Flood Forecasting and Decision Making in the new Millennium. Where are We? **Water Resources Management**, v. 31, n. 10, p. 3111–3129, 2017. Water Resources Management.

VALIPOUR, M.; BANIHABIB, M. E.; BEHBAHANI, S. M. R. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. **Journal of Hydrology**, v. 476, p.433-441, 2013.

WASZCZYSZYN, Z. Fundamentals of Artificial Neural Networks. **Neural Networks in The Analysis and Design of Structures**. p.1–51, 1999. Berlin.

YASEEN, Z. M.; EL-SHAFIE, A.; JAAFAR, O.; AFAN, H. A.; SAYL, K. N. Artificial intelligence based models for stream-flow forecasting: 2000-2015. **Journal of Hydrology**, v. 530, p. 829–844, 2015. Elsevier B.V. Available at: <<http://dx.doi.org/10.1016/j.jhydrol.2015.10.038>>.

ZADEH, M. R.; AMIN, S.; KHALILI, D.; SINGH, V. P. Daily Outflow Prediction by Multi Layer Perceptron with Logistic Sigmoid and Tangent Sigmoid Activation Functions. **Water Resources Management**, v. 24, n. 11, p. 2673–2688, 2010.

ZHANG, D.; LINDHOLM, G.; RATNAWEERA, H. Use long short-term memory to enhance Internet of Things for combined sewer overflow monitoring. **Journal of**

**Hydrology**, v. 556, p. 409–418, 2018. Elsevier B.V. Available at:  
<<https://doi.org/10.1016/j.jhydrol.2017.11.018>>

## APPENDIX

### A.1 AUTOCORRELATION AND PARTIAL AUTOCORRELATION FUNCTIONS

From Figure 20 to Figure 25 the autocorrelation and partial autocorrelation functions for all the series are presented. The blue lines represent the significance limit.

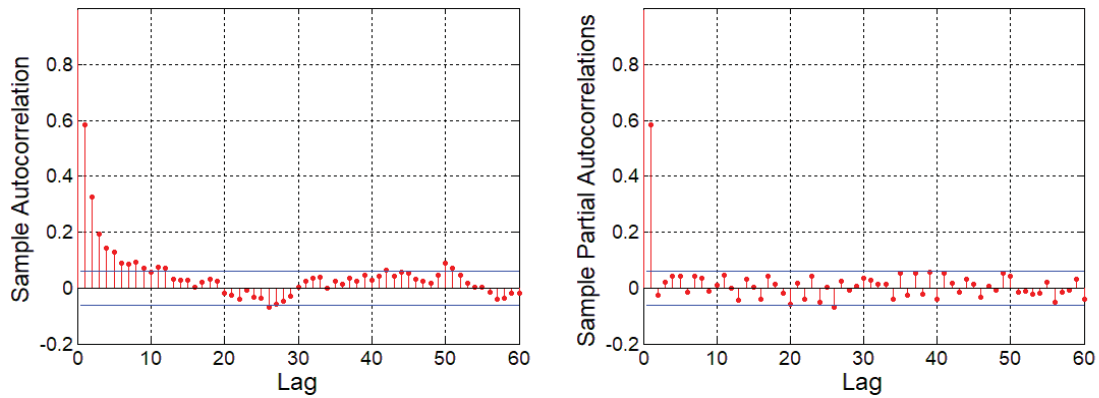


Figure 20. ACF and PACF at Foz do Areia

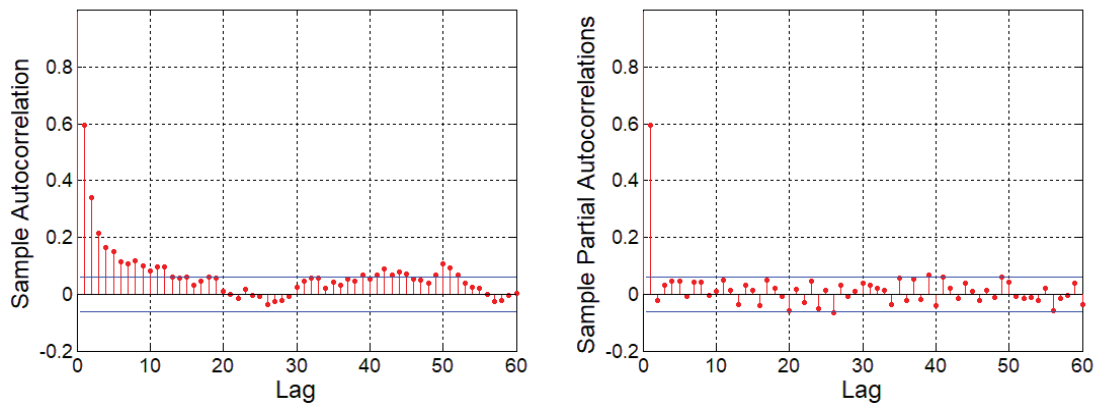


Figure 21. ACF and PACF at Segredo

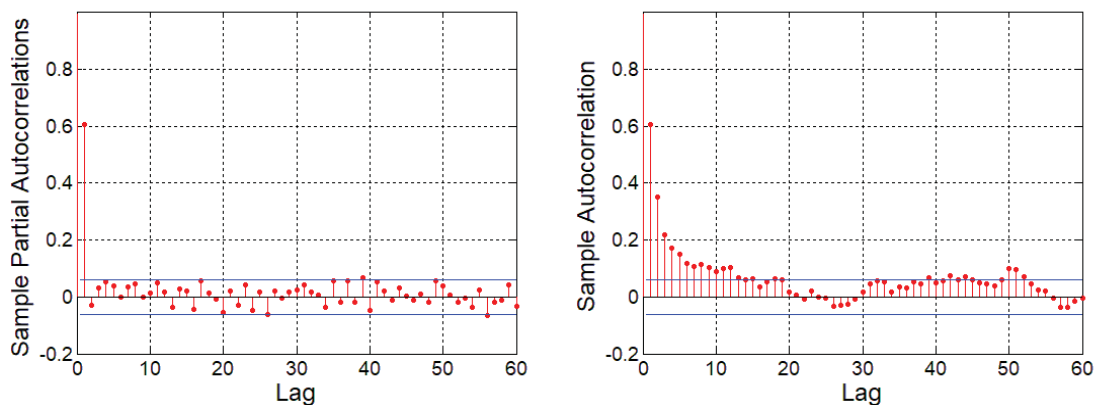


Figure 22. ACF and PACF at Salto Santiago

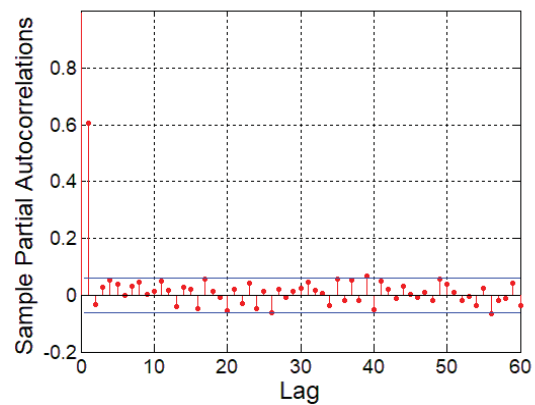
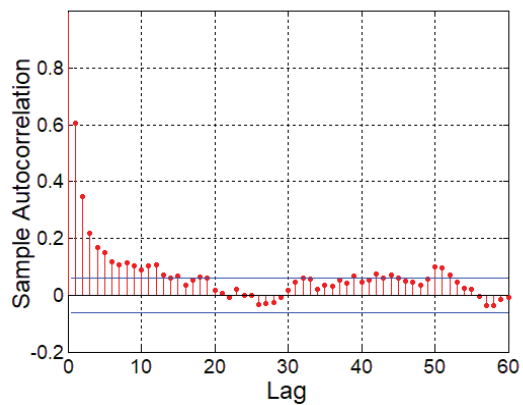


Figure 23. ACF and PACF at Salto Osório

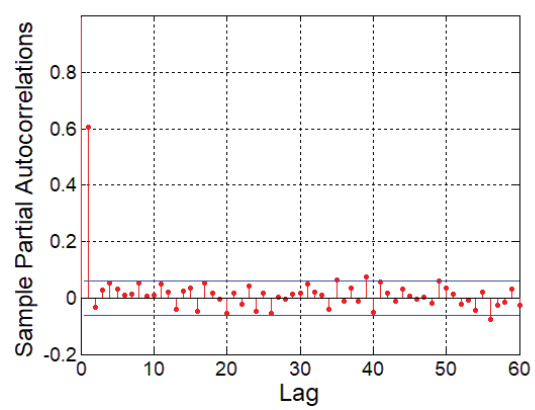
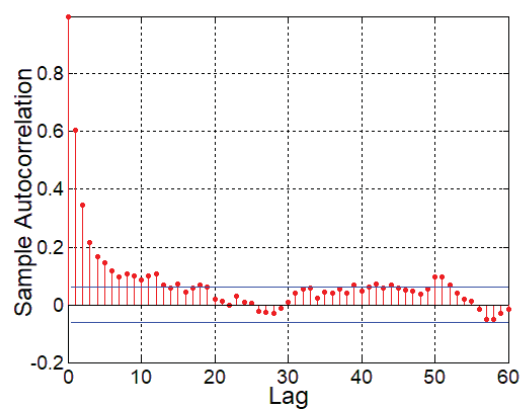


Figure 24. ACF and PACF at Gov. José Richa

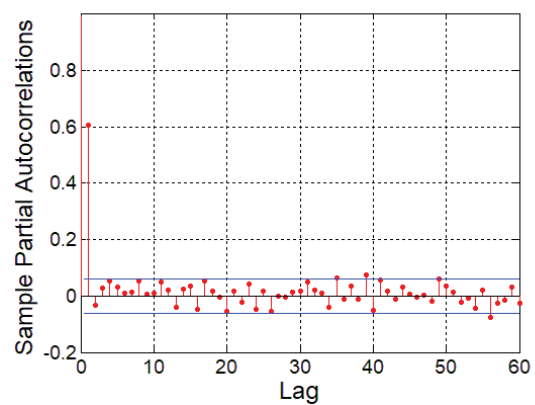
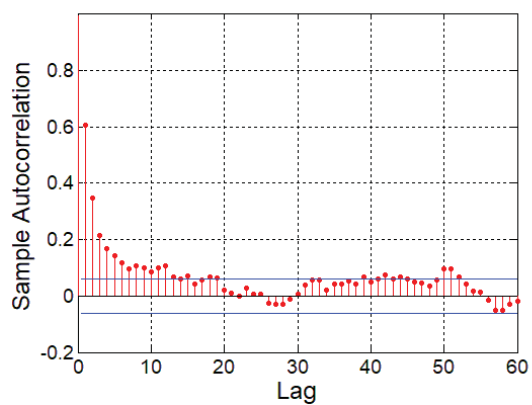


Figure 25. ACF and PACF at Baixo Iguaçu

## A.2 HISTOGRAM PLOTS - RESIDUALS

Figure 26 presents the histogram plots for the ARIMA (1,0,0) residuals at the six stations.

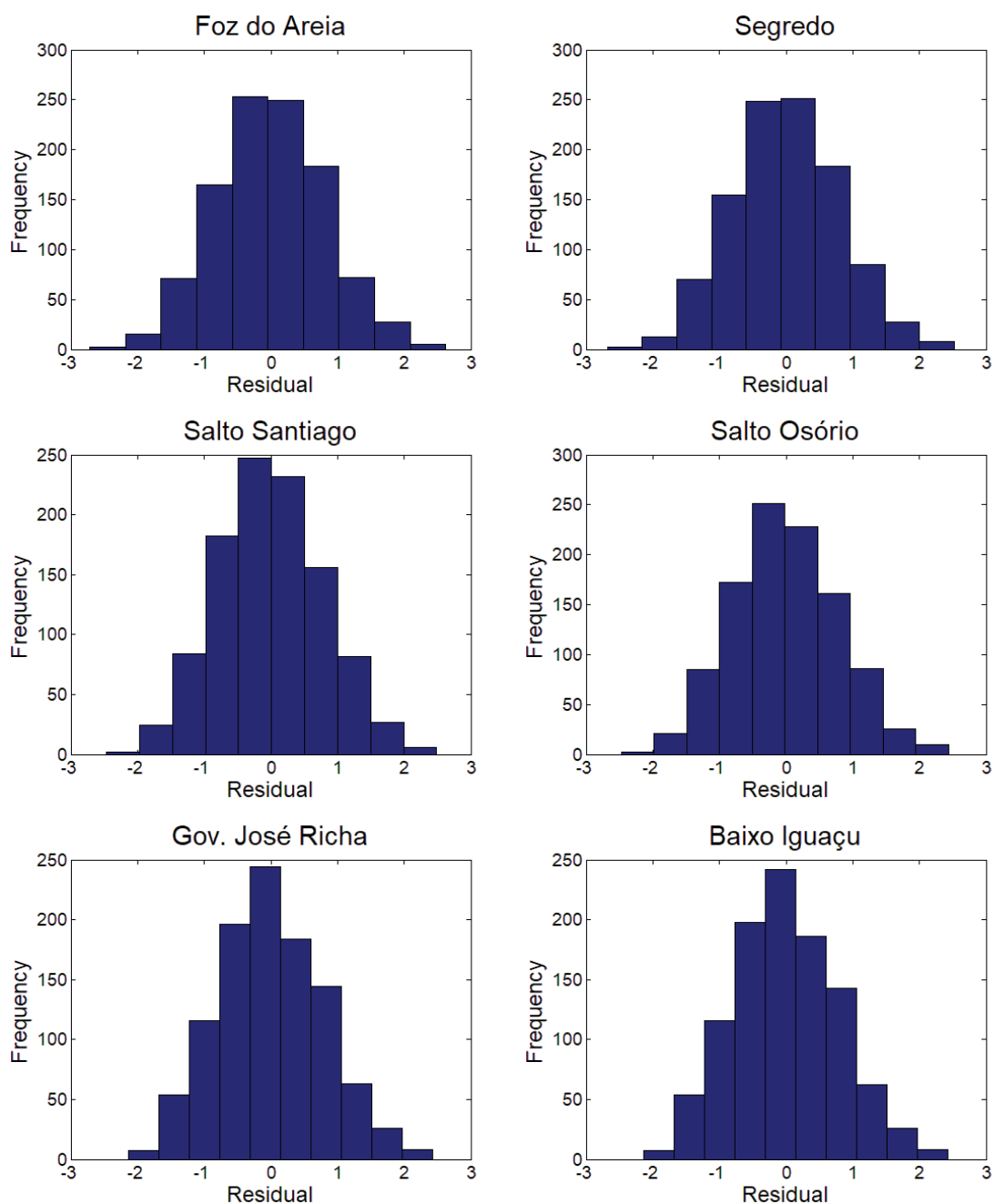


Figure 26. Residual series histograms



### A.3 MONTHLY AVERAGE AND STANDARD DEVIATION

From Figure 27 to Figure 32 the monthly average and standard deviation are presented. Bars indicate means, and lines indicate standard deviations.

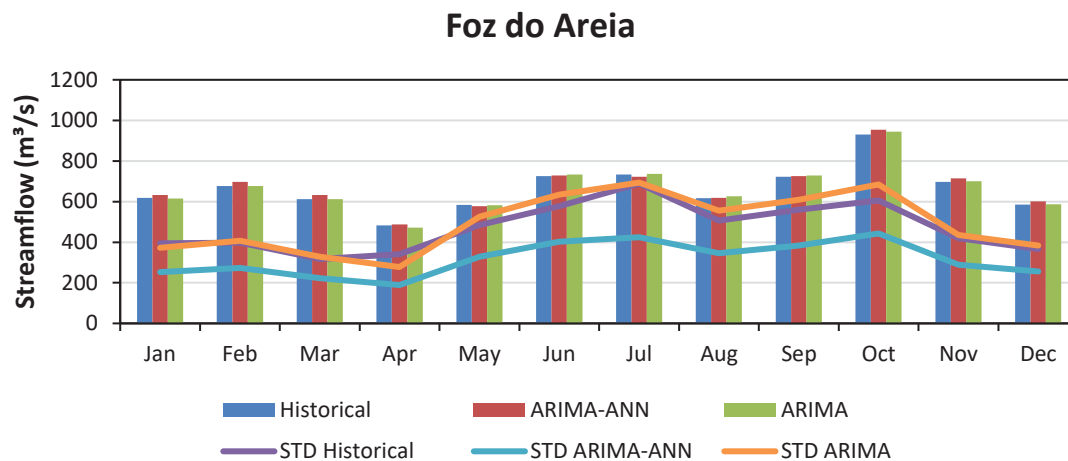


Figure 27. Monthly statistics – Foz do Areia.

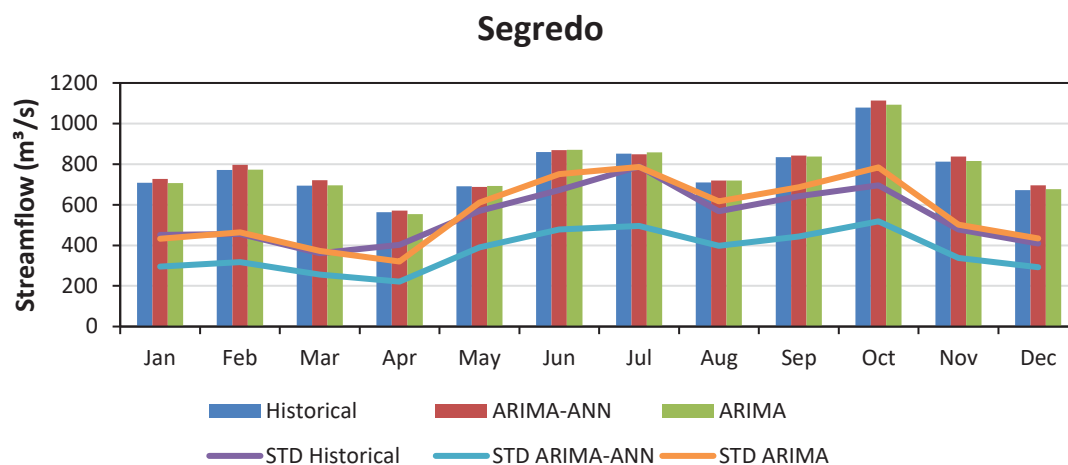


Figure 28. Monthly statistics – Segredo.

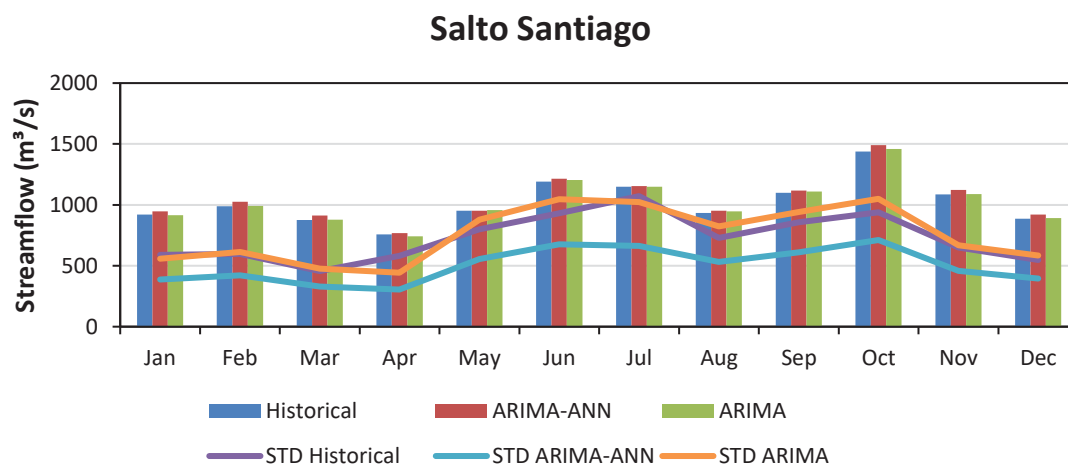


Figure 29. Monthly statistics – Salto Santiago.

### Salto Osório

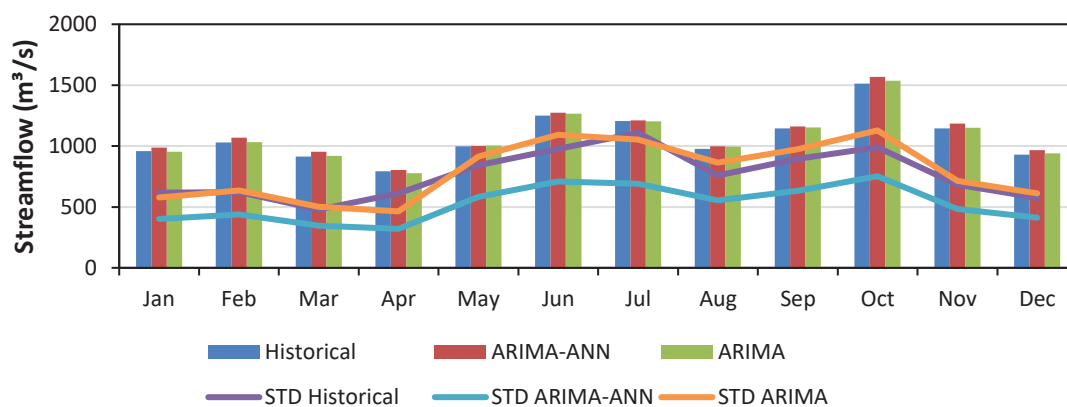


Figure 30. Monthly statistics – Salto Osório.

### Gov. José Richa

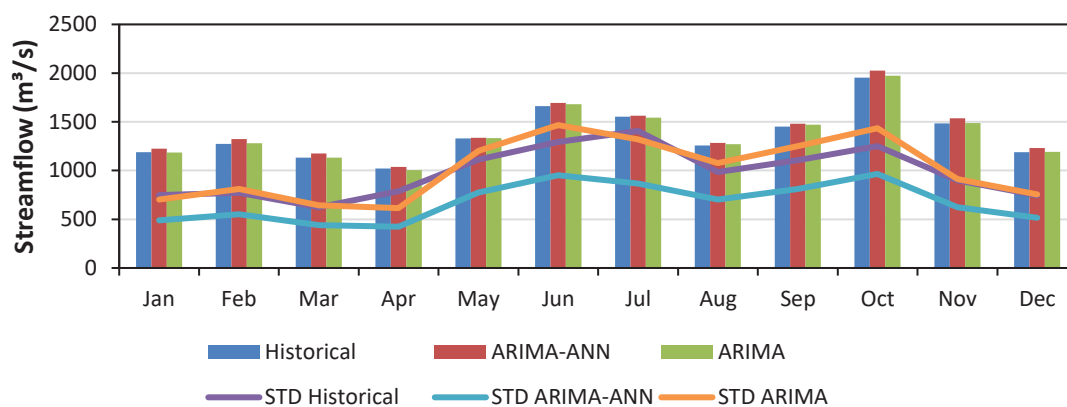


Figure 31. Monthly statistics – Gov. José Richa.

### Baixo Iguaçu

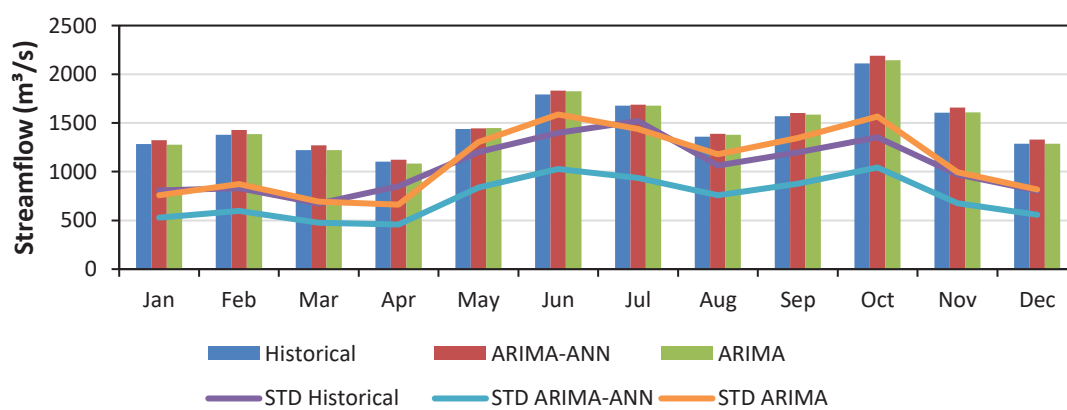


Figure 32. Monthly statistics – Baixo Iguaçu.

#### A.4 AUTOCORRELATION FUNCTION COMPARISON

Figure 33 shows the autocorrelation functions (ACF) for the Historical series at all stations, in contrast with the average ACFs for ARIMA-ANN and single ARIMA synthetic series.

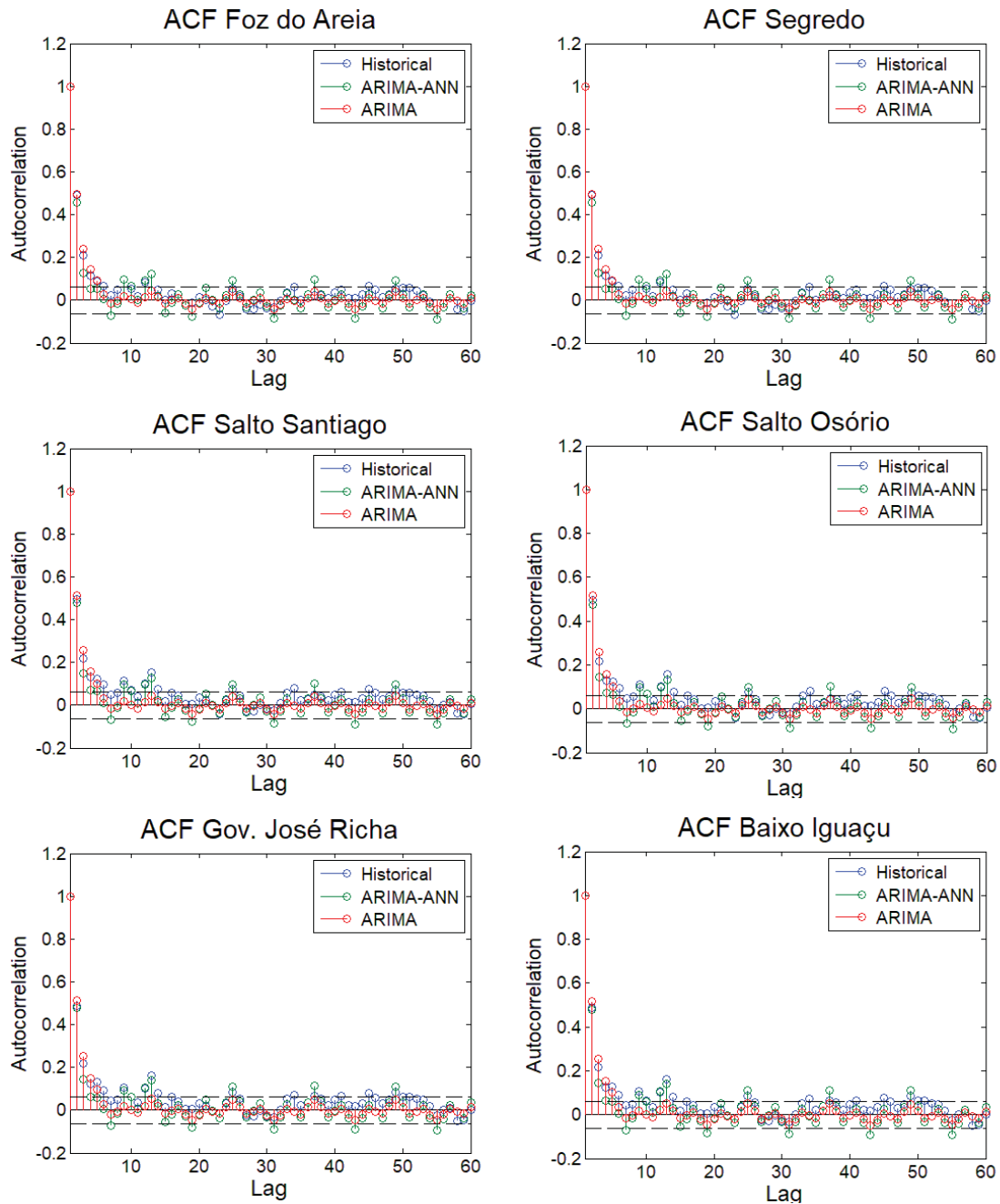


Figure 33. Autocorrelation Functions – Foz do Areia